

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
1 August 2002 (01.08.2002)

PCT

(10) International Publication Number  
**WO 02/059297 A2**

(51) International Patent Classification<sup>7</sup>: C12N 15/10

(21) International Application Number: PCT/DK02/00056

(22) International Filing Date: 25 January 2002 (25.01.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
PA 2001 00128 25 January 2001 (25.01.2001) DK  
PA 2001 00679 1 May 2001 (01.05.2001) DK  
60/300,863 27 June 2001 (27.06.2001) US

(71) Applicant (*for all designated States except US*): **EVOLVA BIOTECH A/S** [DK/DK]; Fruebjergvej 3, DK-2100 Copenhagen Ø (DK).

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **GOLDSMITH, Neil** [GB/GB]; 17 Southfield Road, Oxford OX4 1NX (GB). **SØRENSEN, Alexandra, M., P., SantAna** [DK/DK]; Østre Paradisvej 5D, st.th., DK-2840 Holte (DK). **NIELSEN, Søren, V., S.** [DK/DK]; Høveltsvangvej 72, DK-3450 Allerød (DK).

(74) Agent: **HØIBERG APS**; Store Kongensgade 59 B, DK-1264 Copenhagen K. (DK).

(81) Designated States (*national*): AE, AG, AL, AM, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), DE (utility model), DK (utility model), DM, DZ, EC, EE (utility model), ES, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: A LIBRARY OF A COLLECTION OF CELLS

(57) Abstract: The present invention relates to combinatorial gene expression libraries and methods for making these. Such libraries are useful in discovery of novel and/or enhanced metabolic pathways leading to the production of novel compounds for e.g. drug discovery and/or to the production of known compounds in novel quantities or in novel compartments of the cells. The expression libraries in particular are composed of host cells capable of co-ordinated and controllable expression of large numbers of heterologous genes in the host cells.



WO 02/059297 A2

### **A library of a collection of cells**

This application is a nonprovisional of U.S. provisional application Serial No. 60/300,863 filed 27. June 2001, which is hereby incorporated by reference in its entirety. The application claims priority from Danish patent application number PA 2001 00128 filed 25. January 2001 and PA 2001 00679 filed on 1. May 2001, which are hereby incorporated by reference in their entirety. All patent and nonpatent references cited in the application, or in the present application, are also hereby incorporated by reference in their entirety.

The present invention relates to a library of a collection of cells and a method for producing said library. The library is useful as a starting material for evolving cells or compositions having new properties.

### **Technical field**

The present invention relates to combinatorial gene expression libraries and methods for making these. Such libraries are useful in discovery of novel and/or enhanced metabolic pathways leading to the production of novel compounds for e.g. drug discovery and/or to the production of known compounds in novel quantities or in novel compartments of the cells.

### **Background of the invention**

Methods are known to provide recombined combinatorial gene expression libraries by crossing and recombination between cells comprising expression constructs (WO 00/52180 Terragen Discovery Ltd). Through the recombination, which may be carried out in vitro using the recA recombination enzyme, novel genes are obtained, which may or may not be functional in the host cell.

One drawback of the libraries of the prior art is that evolution of the libraries may only be obtained through crossing and recombination between cells whereby homologous or homeologous genes are recombined thereby resulting in novel genes yielding gene products with slightly changed properties such as substrate specificity, solubility, cellular location etc.

Furthermore once the expression constructs have been inserted into the cells the specific gene combinations of a cell is static. Novel combinations may be obtained by crossing and recombination, but this will also lead to formation of novel genes through cross-over. The novel genes may or may not be functional anymore.

Furthermore, the expression of the inserted expression construct is a co-expression of all the genes inserted into any one cell. When a large number of heterologous genes from a wide variety of distantly related species is assembled in one cell, chances are great that some of the heterologous genes are lethal or sub-lethal to the cell, or that several gene products will compete for the same substrates. When only co-expression of the inserts is possible novel metabolic pathways may remain undiscovered due to this fact or due to the fact that the novel metabolite was being further metabolised to a known metabolite by another inserted enzyme.

### Summary of the invention

According to a first aspect the invention relates to a library comprising a collection of individual cells, the cells being denoted

cell<sub>1</sub>, cell<sub>2</sub>, ..., cell<sub>i</sub>, wherein  $i \geq 2$ ,  
each cell comprising at least one concatemer of individual oligonucleotide cassettes, each concatemer comprising a nucleotide sequence of the following formula:

$$[rs_2\text{-}SP\text{-}PR\text{-}X\text{-}TR\text{-}SP\text{-}rs_1]_n$$

wherein  $rs_1$  and  $rs_2$  together denote a restriction site, SP denotes a spacer of at least two bases, X denotes an expressible nucleotide sequence, PR denotes a promoter, capable of regulating the expression of X in the cell, TR denotes a terminator, and  $n \geq 2$ , and

wherein at least one concatemer of cell<sub>1</sub> is different from a concatemer of cell<sub>2</sub>.

The library according to this embodiment of the invention may in any one cell comprise a unique and preferably random combination of a high number of expression cassettes being heterologous to the host cells. Through this random combination of expression cassettes novel and unique combinations of gene products are obtained in each cell. Such libraries are especially adapted in the

discovery of novel metabolic pathways created through the non-native combinations of gene products.

5 Due to the common structure of the expression cassettes, these may be assembled easily into concatemers and inserted into the host cells via appropriate vectors. Furthermore, the cassettes may at any point be excised from the host cells again using a restriction enzyme specific for the  $rs_1$ - $rs_2$  restriction site preferably without excising the host cell's native genes. After excision the expression cassettes may be mixed with other expression cassettes of similar structure and be re-concatenated  
10 and re-inserted into another host cell in another combination.

A further advantage of the common structure of the expression cassettes, is that the common  $rs_1$ - $rs_2$  sequence may be used as a tag for targeted PCR amplification of the expression constructs.  
15

The expressible nucleotide sequences may conveniently arise from a cDNA library obtained from one or more expression states, wherein the cDNA clones have been inserted into expression constructs. Following excision of the expression construct from the vector comprising the construct in the cDNA library, the multitude of  
20 constructs may be concatenated and inserted into a host cell.

Each unique cell according to the invention may comprise a selection of expressible nucleotide sequences from just one expression state and can thus be assembled from one library representing this expression state or it may comprise cassettes  
25 from of a number of different expression states. The variation among and between cassettes in the cells may be such as to minimise the chance of cross over as the host cell undergoes cell division such as through minimising the level of repeat sequences occurring in concatemers, since it is not an object of this embodiment of the invention to obtain inter- or intrachromosomal recombination of the concatemers.  
30 Nor to obtain recombination with epitopes of the host cell.

The contents of the concatemers may be mixed according to any criteria. Thus a library or a sub-library of individual cells may comprise cells having a common phenotype, cells comprising expression cassettes from a common source, cells  
35 comprising specific combinations of promoter and expressible nucleotide

sequences. A library or sub-library may also or alternatively comprise a collection of individual cells comprising one or more common concatemers in addition to differing concatemers, wherein the common concatemer may represent expression constructs from a common source or coding for genes with a property in common.

5

According to another aspect the invention relates to a library comprising a collection of individual cells, the cells being denoted

cell<sub>1</sub>, cell<sub>2</sub>, ..., cell<sub>i</sub>, wherein  $i \geq 2$ ,

each cell comprising at least two expression cassettes comprising a nucleotide sequence of the following formula:

10

[rs<sub>2</sub>-SP-PR-X-TR-SP-rs<sub>1</sub>]

wherein rs<sub>1</sub> and rs<sub>2</sub> together denote a restriction site, SP denotes a spacer of at least two bases, X denotes an expressible nucleotide sequence, PR denotes a promoter, capable of regulating the expression of X in cell, TR denotes a terminator, and

15

wherein at least one of the expression cassettes comprises an expressible nucleotide sequence heterologous to the to cell, and at least one of the cassettes of cell<sub>1</sub> is different from the cassettes of cell<sub>2</sub>.

20

According to this aspect of the invention, the cells are defined with reference to the expression cassettes. This aspect of the invention shares many advantages with the first aspect of the invention.

25

According to a third aspect the invention relates to a library comprising a collection of individual cells, the cells being denoted

cell<sub>1</sub>, cell<sub>2</sub>, ..., cell<sub>i</sub>, wherein  $i \geq 2$ ,

each cell comprising a random combination of heterologous oligonucleotides having the general formula:

[PR-X]

30

wherein X denotes an expressible nucleotide sequence, and PR denotes an independently controllable promoter being operably associated with X.

35

In a library according to this aspect of the invention, the mixing of gene products may not only be done upon insertion of the expressible nucleotide sequences, but also during expression by inducing and/or repressing one or more promoters each

regulating the expression of a random group of expressible nucleotide sequences. Thus in each cell, a unique sub-set of genes may be induced and/or repressed at any point

5 This feature adds another level of potential variation in the discovery of novel biochemical pathways. By the up and down regulation of independent promoters any combination of sub-sets of genes may be turned on or off in a population of cells having a random combination of promoters and expressible nucleotide sequences.

10 In the evolution of novel biochemical pathways based on the insertion and expression of a high number of heterologous genes in a population of cells, it is highly likely that cells will be killed due to the formation of lethal gene products. If each cell comprises just one lethal gene, the co-expression of a number of heterologous genes will not lead to any novel biochemical pathways. By having a  
15 random combination of promoters and expressible nucleotide sequences, it may be possible to down regulate lethal or sub-lethal genes without affecting the expression of the other heterologous expression constructs.

It is also possible to use the co-ordinated expression obtained through the random  
20 combination of promoters and expressible nucleotide sequences from the same pool of expressible nucleotide sequences to identify expressible nucleotide sequences involved in a desired or unwanted property (e.g. lethality or sub-lethality). In a population according to this aspect of the invention, each cell may in principle comprise more or less the same heterologous expressible nucleotide sequences,  
25 the difference between the cells being the groups of expressible nucleotide sequences that are induced/repressed by a given promoter. In such a population of cells a desired or unwanted property will be identified in different cells following induction/repression of different promoters. As an illustrative example, in cell A the property may be associated with induction of promoter 1, 2, and 3, and in cell B the  
30 property may be associated with induction of promoter 5 and 6. With this information it is possible to target the property (or properties) to the group of expressible nucleotide sequences associated with these promoters in these cells. The expression constructs may be isolated using knowledge about the promoter nucleotide sequence and sequences common for the identified cells may be

identified. Thus, by turning on and off only certain sub-sets of genes at a time, it is possible to identify which gene combinations have given a particular phenotype.

According to a further aspect the invention relates to a library comprising at least one library or at least one sub-library as defined above. In the evolution of novel biochemical pathways, it may be preferable to use a number of libraries or sub-libraries and to evolve these in parallel or mix the libraries in order to improve the chances of identifying a desired property.

According to a further aspect the invention relates to a method of producing a library comprising a collection of individual cells, comprising the steps:

- i) providing a population of nucleotide cassettes having the general formula  $[rs_2\text{-}SP\text{-}PR\text{-}X\text{-}TR\text{-}SP\text{-}rs_1]$ , wherein  $rs_1$  and  $rs_2$  together denote a restriction site, SP denotes a spacer of at least two bases, X denotes an expressible nucleotide sequence, PR denotes a promoter, capable of regulating the expression of X in the cell, TR denotes a terminator, and
- ii) assembling random sub-sets of the cassettes into concatemers comprising at least two cassettes,
- iii) ligating the concatemers into vectors,
- iv) introducing vectors into host cells,
- v) mixing at least two cells so that at least one concatemer of a first cell comprises a random sub-set of cassettes being different from a random sub-set of cassettes of a concatemer of a second cell.

The assembly of concatemers is facilitated by the common structure of the expression cassettes. When the  $rs_1$ - $rs_2$  restriction site produces sticky ends with a predetermined nucleotide sequence the assembly of the concatemers becomes especially easy to perform.

The randomisation of the cassettes may be done at any stage, i.e. during a preceding step in which an entry library (for storing and amplifying cassettes) is produced or during the insertion into vectors and/or during the transformation into host cells. Preferably the randomisation is done during the concatenation step.

According to another aspect the invention relates to a method of producing a library comprising a collection of individual cells, comprising the steps:

- i) inserting at least two expressible nucleotides into the cloning site of at least two primary vectors comprising a cassette, the cassette comprising a nucleotide sequence of the general formula in 5'→3' direction: [RS1-RS2-SP-PR-CS-TR-SP-RS2-RS1'] wherein RS1 and RS1' denote first restriction sites, RS2 denotes another restriction site different from RS1 and RS1', SP denotes a spacer sequence of at least two nucleotides, PR denotes a promoter, CS denotes a cloning site, and TR denotes a terminator.
- ii) excising the cassettes using at least a restriction enzyme specific for RS1, RS1' and RS2 obtaining expression cassettes having the general formula [rs<sub>2</sub>-SP-PR-X-TR-SP-rs<sub>1</sub>], wherein rs<sub>1</sub>-rs<sub>2</sub> together denote a restriction site, and wherein X denotes an expressible nucleotide sequence,
- iii) inserting the expression cassettes into a vector,
- iv) transferring the expression cassettes into at least two host cells, and
- v) mixing at least two host cells having different cassettes.

According to this method for producing a library of individual cells the source expressible nucleotide sequences are first ligated into a primary vector comprising a cloning site and a cloning cassette. This primary vector may be maintained in a cDNA library and reisolated for excision of the expression cassettes and insertion into a host cell. Through this process the expressible nucleotide sequences are given a common structure which makes it possible to clone the cassettes into a predetermined cloning site in a vector and to remove the cassettes selectively from the host cells later.

According to a final aspect the invention relates to a method of producing a library comprising a collection of individual cells, comprising the steps:

- i) providing at least one expressible nucleotide sequence,
- ii) ligating at least one expressible nucleotide sequence to a controllable promoter capable of functioning in a host cell obtaining a first expression construct,



- iii) ligating at least one expressible nucleotide sequence to another independently controllable promoter capable of functioning in a host cell, obtaining a second expression construct,
- iv) inserting constructs of step ii) and iii) into at least two host cells,
- 5 v) mixing at least two cells having a different combination of independently controllable promoter and expressible nucleotide sequences.

10 According to this aspect of the invention there is provided a convenient method for preparation of a library of individual cells comprising expressible nucleotide sequences under the operable control of at least two controllable promoters.

#### **Brief description of the drawings**

15 Fig. 1 shows a flow chart of the steps leading from an expression state to incorporation of the expressible nucleotide sequences in an entry library (a nucleotide library according to the invention).

20 Fig. 2 shows a flow chart of the steps leading from an entry library comprising expressible nucleotide sequences to evolvable artificial chromosomes (EVAC) transformed into an appropriate host cell. Fig. 2a shows one way of producing the EVACs which includes concatenation, size selection and insertion into an artificial chromosome vector. Fig. 2b shows a one step procedure for concatenation and ligation of vector arms to obtain EVACs.

25 Fig. 3 shows a model entry vector. MCS is a multi cloning site for inserting expressible nucleotide sequences. Amp R is the gene for ampicillin resistance. Col E is the origin of replication in E. coli. R1 and R2 are restriction enzyme recognition sites.

30 Fig. 4 shows an example of an entry vector according to the invention, EVE4. MET25 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer 1 and spacer 2 are constituted by a few nucleotides deriving from the multiple cloning site, MCS, SrfI and Ascl are restriction

enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 1.

Fig 5 shows an example of an entry vector according to the invention, EVE5. CUP1 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer 1 and spacer 2 are constituted by a few nucleotides deriving from the multiple cloning site, MCS, SrfI and Ascl are restriction enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 2.

Fig 6 shows an example of an entry vector according to the invention, EVE8. CUP1 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer3 is a 550 bp fragment of lambda phage DNA. Spacer4 is a ARS1 sequence from yeast. SrfI and Ascl are restriction enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 3.

Fig. 7 shows a vector (pYAC4-Ascl) for providing arms for an evolvable artificial chromosome (EVAC) into which a concatemer according to the invention can be cloned. TRP1, URA3, and HIS3 are yeast auxotrophic marker genes, and AmpR is an E. coli antibiotic marker gene. CEN4 is a centromere and TEL are telomeres. ARS1 and PMB1 allow replication in yeast and E. coli respectively. BamH I and Asc I are restriction enzyme recognition sites. The nucleotide sequence of the vector is set forth in SEQ ID NO 4.

Fig 8. shows the general concatenation strategy. On the left is shown a circular entry vector with restriction sites, spacers, promoter, expressible nucleotide sequence and terminator. These are excised and ligated randomly.

Lane	F/Y
1	100/1
2	50/1
3	20/1
4	10/1
5	5/1
6	2/1
7	1/1
8	1/2
9	1/5

Legend: Lane M: molecular weight marker,  $\lambda$ -phage DNA digested w. Pst1. Lanes 1-9, concatenation reactions. Ratio of fragments to yac-arms(F/Y) as in table.

5 Fig 9a and 9b. illustrates the integration of concatenation with synthesis of evolvable artificial chromosomes and how concatemer size can be controlled by controlling the ratio of vector arms to expression cassettes, as described in example 7.

10 Fig 10. Library of EVAC transformed population shown under 4 different growth conditions. Coloured phenotypes can be readily detected upon induction of the Met25 and/or the Cap1 promoters.

Fig 11. EVAC gel Legend: PFGE of EVAC containing clones :  
Lanes. a: Yeast DNA PFGE markers(strain YNN295), b: lambda ladder, c: non-  
15 transformed host yeast, 1 – 9 : EVAC containing clones. EVACs in size range 1400-1600 kb. Lane 2 shows a clone containing 2 EVACs sized ~1500 kb and ~550 kb respectively. The 550kb EVAC is comigrating with the 564kb yeast chromosome and is resulting in an increased intensity of the band at 564 kb relative to the other bands in the lane. Arrows point up to EVAC bands.

20

### Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of skill in the art to which this  
25 invention belongs.

As used herein, growth under selective conditions, means growth of a cell under conditions that require expression of a selectable marker for survival.

30 By a controllable promoter is meant a promoter, which can be controlled through external manipulations such as addition or removal of a compound from the surroundings of the cell, change of physical conditions, etc.

35 An independently controllable promoter may be induced/repressed substantially without affecting the induction/repression of other promoters according to the

invention. The induction/repression of an independently controllable promoter may affect native promoters in the host cells.

Co-ordinated expression refers to the expression of a sub-set of genes which are induced or repressed by the same external stimulus.

#### Oligonucleotides

Any fragment of nucleic acids having approximately from 2 to 10000 nucleic acids.

#### Restriction site

For the purposes of the present invention the abbreviation RSn (n=1,2,3, etc) is used to designate a nucleotide sequence comprising a restriction site. A restriction site is defined by a recognition sequence and a cleavage site. The cleavage site may be located within or outside the recognition sequence. The abbreviation "rs<sub>1</sub>" or "rs<sub>2</sub>" is used to designate the two ends of a restriction site after cleavage. The sequence "rs<sub>1</sub>-rs<sub>2</sub>" together designate a complete restriction site.

The cleavage site of a restriction site may leave a double stranded polynucleotide sequence with either blunt or sticky ends. Thus, "rs<sub>1</sub>" or "rs<sub>2</sub>" may designate either a blunt or a sticky end.

In the notation used throughout the present invention, formulae like:

RS1-RS2-SP-PR-X-TR-SP-RS2-RS1

should be interpreted to mean that the individual sequences follow in the order specified. This does not exclude that part of the recognition sequence of e.g. RS2 overlap with the spacer sequence, but it is a strict requirement that all the items except RS1 and RS1' are functional and remain functional after cleavage and re-assembly. Furthermore the formulae do not exclude the possibility of having additional sequences inserted between the listed items. For example introns can be inserted as described in the invention below and further spacer sequences can be inserted between RS1 and RS2 and between TR and RS2. Important is that the sequences remain functional.

Furthermore, when reference is made to the size of the restriction site and/or to specific bases within it, only the bases in the recognition sequence are referred to.

#### Expression state

An expression state is a state in any specific tissue of any individual organism at any one time. Any change in conditions leading to changes in gene expression leads to another expression state. Different expression states are found in different individuals, in different species but they may also be found in different organs in the same species or individual, and in different tissue types in the same species or individual. Different expression states may also be obtained in the same organ or tissue in any one species or individual by exposing the tissues or organs to different environmental conditions comprising but not limited to changes in age, disease, infection, drought, humidity, salinity, exposure to xenobiotics, physiological effectors, temperature, pressure, pH, light, gaseous environment, chemicals such as toxins.

#### Artificial chromosome

As used herein, an artificial chromosome (AC) is a piece of DNA that can stably replicate and segregate alongside endogenous chromosomes. For eukaryotes the artificial chromosome may also be described as a nucleotide sequence of substantial length comprising a functional centromere, functional telomeres, and at least one autonomous replicating sequence. It has the capacity to accommodate and express heterologous genes inserted therein. It is referred to as a mammalian artificial chromosome (MAC) when it contains an active mammalian centromere. Plant artificial chromosome and insect artificial chromosome (BUGAC) refer to chromosomes that include plant and insect centromeres, respectively. A human artificial chromosome (HAC) refers to a chromosome that includes human centromeres, AVACs refer to avian artificial chromosomes. A yeast artificial chromosome (YAC) refers to chromosomes that are functional in yeast, such as chromosomes that include a yeast centromere.

As used herein, stable maintenance of chromosomes occurs when at least about 85%, preferably 90%, more preferably 95% of the cells retain the chromosome. Stability is measured in the presence of a selective agent. Preferably these chromosomes are also maintained in the absence of a selective agent. Stable chromosomes also retain their structure during cell culturing, suffering neither intrachromosomal nor interchromosomal rearrangements.

**Detailed description of the invention**

The present invention relates to libraries of individual cells useful for capturing and preserving a diversity of genetic resources from nature, and for expressing the captured genetic resources and allowing them to interact to produce a diversity of chemical structures. The invention also facilitates screening for desirable properties and compounds.

More particularly, the invention provides methods for constructing and screening libraries of individual cells comprising heterologous expressible nucleotide sequences. These libraries comprise random assortments of expressible nucleotide sequences from multiple expression states and preferably also from multiple species the products of which are allowed to interact with each other in the expression host, and result in some cases in the formation of novel biochemical pathways and/or the production of novel classes of compounds. Moreover, the libraries of the invention provide efficient access to otherwise inaccessible sources of molecular diversity.

The novel biochemical pathways may carry out processes including but not limited to structural modification of a compound, addition of chemical groups to the compound, or decomposition of the compound.

The novel classes of compound may include but are not limited to metabolites, secondary metabolites, enzymes, or structural components of an organism. A compound of interest may have one or more potential therapeutic properties, including but not limited to agonist or antagonist to a class of receptor or a particular receptor, antibiotic, antiviral, antitumor, pharmacological or immunomodulating properties or be other commercially-valuable chemicals such as pigments.

A library of individual cells is a library comprising expression constructs prepared from randomly assembled or even concatenated expressible nucleotide sequences derived from a plurality of species of donor organisms, in which expressible nucleotide sequences are operably associated with regulatory regions that drives expression of the expressible nucleotide sequences in an appropriate host organism. The host organisms used are capable of producing functional gene

products of the donor organisms. Upon expression in the host organism, gene products of the donor organism(s) may interact to form novel biochemical pathways.

5 Generally, the methods of the invention comprise providing expressible nucleotide sequences derived from one or more donor organism(s), engineering said expressible nucleotide sequences into a context where said expressible nucleotide sequences can be transcribed in a given host organism, and introducing said expressible nucleotide sequences into a host organism via a cloning or expression vector so that one or more expressible nucleotide sequences of the donor  
10 organism(s) are transferred to and expressed in the host organism. Such host organisms containing donor expressible nucleotide sequences are pooled to form a library.

The transferred genetic material, typically comprises a random assortment of  
15 expressible nucleotide sequences, the expression of which is driven and controlled by one or preferably by more functional regulatory regions. The expression construct or vector advantageously provide these regulatory regions. The expressible nucleotide sequences of the donor organism(s) are transcribed, translated and processed in the host organism to produce functional proteins that in turn generate  
20 the metabolites of interest.

Once a desirable activity or compound is identified, downstream drug development efforts such as strain improvement and process development, are greatly facilitated. The positive clone can be cultured under standard conditions to produce the desired  
25 compound in substantial amounts for further studies or uses. The expressible nucleotide sequences of the biochemical pathway are immediately available for sequencing, mutation, expression, and further rounds of screening. The cloned biochemical pathway is readily amenable to traditional and/or genetic manipulations for overproduction of the desired compound.

30

Furthermore, according to the embodiments comprising the expression cassettes with common structure, several positive cells may be identified, their expression cassettes be excised due to the presence of a common restriction site, which is preferably a rare restriction site. The excised expression cassettes may be re-

assembled in a random or targeted manner to produce novel combinations of the selected expression cassettes.

5 Furthermore, biochemical pathways that are otherwise silent or undetectable in the donor organism may be discovered more easily by virtue of their functional reconstitution in the host organism. Since the biochemical characteristics of the host organism are well known, many deviations as a result of expression of donor genetic material can readily be recognised. Novel compounds may be detected by comparing extracts of a host organism containing donor genetic material against a  
10 profile of compounds known to be produced by the control host organism under a given set of environmental conditions. Even very low levels of a desirable activity or compound may be detected when the host biochemical and cellular background of the host organism is well characterised.

15 In one embodiment, the methods may be applied to donor organism(s) that cannot be recovered in substantial amounts in nature, or cultured in the laboratory. By transferring genetic material such as cDNA from such organisms into a host organism, the organisms' metabolic pathways may be reproduced, and their products tested efficiently for any desirable properties. Thus, the genetic diversity of  
20 these organisms is captured and preserved and combined with the genetic diversity of other organisms.

In another embodiment of the invention, a library can be constructed in which the expressible nucleotide sequences from one or multiple donor organisms are  
25 randomly concatenated prior to introduction into the host organism. Thus, each host organism in the library may individually contain a unique, random combination of expressible nucleotide sequences derived from the various donor pathways or organisms. For the most part, such combinations of expressible nucleotide sequences in the library do not occur in nature. Upon expression, the functional  
30 gene products of the various donor pathways or organisms interact with each other and with the native host complement of gene products in individual host organisms to generate combinations of biochemical reactions which result in novel metabolic pathways and/or production of novel compounds. Collectively, the genetic resources of the donor organisms in the library are translated into a diversity of chemical  
35 compounds that may not be found in individual donor organisms.



5 In another aspect of the invention, the methods may be applied to the generation of a multiple kingdom pathway in the host organism. An example of this would be the introduction of genes from carotenoid pathways (obtained from fungi, algae and/or plants) as well as genes from synthesis of Vitamin A (obtained from animals) or genes coding for the production of visual pigments (obtained from insects). By such targeted selection and combination of elements of biochemical pathways across kingdoms the likelihood of obtaining novel metabolites may be further increased.

10 In another aspect of the invention, the species of donor organisms may be selected on the basis of their biological characteristics. Such biological characteristics may include, but are not limited to the capability to utilise certain nutrients, to survive under extreme conditions, to derivatise a chemical structure, and the ability to break down or catalyse formation of certain types of chemical linkages. When expressible  
15 nucleotide sequences of the donor organism are expressed in the host organism, the donor gene products can modify and/or substitute the functions of host gene products that constitute host metabolic pathways, thereby generating novel hybrid pathways. Novel activities and/or compounds may be produced by hybrid pathways comprising donor and host-derived components. The target metabolic pathway  
20 modified by donor gene products may be native to the host organism. Alternatively, the target metabolic pathway may be provided by products of heterologous genes which are endogenous or have been genetically engineered into every host organism prior to or contemporaneous to construction of the gene expression library. Thus, the present invention also embodies constructing and screening gene  
25 expression libraries, wherein DNA fragments encoding metabolic pathway of donor organisms are cloned and coexpressed in host organisms containing a target metabolic pathway.

30 In another embodiment of the invention, the host organism may have an enhanced complement of active drug efflux systems which secretes the compounds of interest into the culture medium, thus reducing the toxicity of the compounds to the host organism. Absorptive material, e.g., neutral resins, may be used during culturing of the host organisms, whereby metabolites produced and secreted by the host organism may be sequestered, thus facilitating recovery of the metabolites.

In many respects, the libraries provides significant convenience and time advantage to the various steps of development of novel small molecules such as the development of drugs up to clinical trials. The libraries of the invention are compatible with e.g. the established multi-well footprint format and robotics for high-throughput screening. The host organisms of the invention are organisms commonly used for genetic manipulation and/or process development. The present invention takes advantage of the fact that such host organisms or production hosts are well-characterised in terms of their biological properties and maintenance requirements. By transferring genetic materials from a donor organism to other more familiar expression systems, the need for difficult culturing conditions for the donor organism is reduced. Thus, the biological activities, the pharmacokinetic and toxic properties of any lead compound discovered in the system of the invention may be studied and optimised more efficiently.

The novel metabolic pathway generated in a positive clone can be delineated by standard techniques in molecular biology. The lead compound may be synthesised by culturing a clone of the drug-producing host organism under standard or empirically determined culture conditions, so that sufficient quantities of the lead compound may be isolated for further analysis and development. There are already high purity manufacturing protocols, such as Good Manufacturing Practice (GMP) established for some of these standard industrial host organisms. Unlike conventional methods of screening natural product sources, less effort is required to adapt the screening and production technologies to the particular requirements of each potentially drug-producing organism.

The present invention also provides libraries made according to the methods of the invention from genetic materials of a particular set of donor organisms and/or cell types. Not all organisms or cell types in a set, especially mixed samples, need to be individually identified or characterised to enable preparation of the libraries.

Any library of the invention may be amplified, replicated, and stored. Amplification is preferably performed by introducing entry vectors containing expressible nucleotide sequences in a initial host organism such as *E. coli* so that so that multiple clones of the expressible nucleotide sequences are produced. Replication refers to picking

and growing of individual clones in the library. A library of the invention may be stored and retrieved by any techniques known in the art that is appropriate for the host organism. Thus, the libraries of the invention are an effective means of capturing and preserving the genetic resources of donor organisms, which may be accessed repeatedly in a drug discovery program or other discovery programs.

### **Concatemer Assemblage**

Concatemers may be assembled from cDNA libraries on a routine basis. A typical concatemer generation step will pool e.g. 1,000 genes = cDNA expression constructs (from 1 sample) and use this to generate 1,000 concatemers, with an average of 25 genes per concatemer. This means on average each gene will be in 25 different concatemers within a pool. One such concatemer "Source Pool" may be generated per source cDNA library. The Source Pools are suitable for storage of the concatemers.

However, the invention is not limited to any specific number of genes in a source pool. Concatemers with approximately 500 genes are easily produced and it is contemplated that this number can be increased even further

The actual numbers depend on the number of different promoters and/or spacers and/or terminators to be incorporated -i.e. if an expression state gives 1000 different cDNAs and these are to be combined with 2 promoters and/or spacers and/or terminators the numbers increase proportionally: 1000 cDNAs = 2000 expression constructs, so if each construct should still be present in 25 concatemers of 25 constructs then the source pool size would be 2000.

Certain Source Pools may in fact be generated on a function rather than species basis. Such a source pool may for example be based on sources known for a specific property, such as carotenoid activity, pharmaceutical properties, chemotaxonomic properties, etc.

### Host library assemblage

Source Pools may be mixed and used to generate host libraries or screening libraries with each host containing multiple concatemers. In selecting which Source Pools to mix one may use knowledge of the source of given libraries, host pathways, the desired focus of particular programmes and success rates of given libraries in particular screens.

If each source library is constructed from 1,000 different genes and assembled into EVACs each containing 25 genes, then for any one given gene, of those EVACS that do contain the gene, 98.8% of them will contain just one copy, 1.2% will contain 2 copies and 0.01% will contain 3 or more copies. Thus for all practical purposes each EVAC can in this situation be regarded as composed of 25 different genes. Should a cell population be created from four such source libraries, then each different gene (assuming no overlap between genes from different sources) will be represented at a frequency of 1 copy per 4,000 genes.

In a cell population where each cell contains four EVACS, generated from a pool of 4 source libraries, then in respect of any one of the source libraries, statistically:

- 0.4% of cells will have all four of their EVACS from this source
- 4.3% of cells will have three out of four EVACS from this source
- 25.5% of cells will have two EVACS from this source
- 38.3% of cells will have just one EVAC from this source
- 31.6% of cells will not contain any EVACS from this source

From these figures the probability of any two-gene combination can be calculated using standard statistical tools

For more focused evolutionary approaches, such as the evolution of novel carotenoids or other known structural classes or metabolite pathways, EVACs can be enriched for enzymes, and homologs or functional analogs of these enzymes, that conduct different stages of the metabolic pathway. Such an approach can lead to significant probabilities that essentially all steps of a given pathway are represented, at least at the transcription level, in a cell. Thus if a 10-step pathway is required, and 50-gene EVACS are constructed randomly from genes encoding for homologous or analogous enzymes to those responsible for each step then any given step

will be encoded in >85% of EVACs between 3-9 times (inclusive) and will be entirely missing in just 0.52% of EVACs. Thus it can be seen that a  $10^8$  member cell population where each cell contains 4 EVACs of 50 genes each, constructed from 4 enzyme encoding gene pools, will contain a large number of cells in which all steps of the potential pathway are represented, in most cases multiple times.

### Sub-libraries

Initial screens are designed to sort host lines into "collections", sub-libraries, based on whether novel activity has been induced by the concatemers, and the type of activity that has been induced. As such initial screens should be reasonably high throughput and should be arbitrary in their selection criteria.

A large number of such screens can be considered. An illustrative example of such screens may include but are not limited to:

- Novel spectral properties
- Induced cytochrome oxidase activity
- Changed size, morphology, stickiness or adhesive properties or lack thereof
- Ability to grow on substrates they cannot normally grow on
- Ability to grow on sublethal substrates
- Ability to grow in the absence of normal essential requirements
- Ability to grow on media comprising one or more inhibitors
- Ability to grow under changed physical conditions, such as temperature, osmolarity, electromagnetic radiation including light of certain wavelengths.
- Ability to grow under magnetic field of certain force.
- Secretion or the lack of it from the cell
- The inhibition or prevention of inhibition of an enzyme
- The activation of a receptor.
- The prevention of an activating molecule binding to a receptor.
- The inhibition or promotion of binding of small molecules or proteins to nucleic acid or peptide sequences.
- The inhibition or promotion of transcription or translation of post translational processing.
- Changes in the transport or localisation of molecules within the cell or within organelles.

- Changes in the DNA content or morphology of the cell.
- The production of small molecules with certain properties that allow their selective isolation (e.g. all the chromatography principles available to the skilled practitioner).
- 5      • The production of small molecules with certain spectroscopic properties (defined broadly to include visible light, microwaves, IR, UV, X-ray, etc.).
- Changes in the morphology of the cell, including the prevention or promotion of cell differentiation.
- The induction of apoptotic pathways.

10

For each Host Library (of 10,000 host lines) the 1-2 % of host lines that are most extreme on each of such criteria may be grouped into a sub-library. These initial sorting screens will in general be conducted under conditions that maximise the number of genes expressed per concatemer.

15

The output of a sorting screen may be host lines that are characterised on one or more broad criteria. These may be categorised as sub-libraries.

20

A sub-library may be defined with reference to a common phenotype of the cells in the sub-library. But a sublibrary may also be defined as a collection of individual cells, said cells having – for at least one identical expressible DNA sequence – different promoters, i.e. with reference to the presence of specific expressible nucleotide sequences. Furthermore, a sub-library may be described with reference to a cassette and/or in a concatemer of cassettes comprised in the host cells. A sub-

25      library may thus be defined as a collection of individual cells, each cell having - in at least one cassette of the concatemer - identical expressible DNA sequences. A sublibrary may also be looked upon as a collection of individual cells, said cells having – for at least one identical expressible DNA sequence, more preferably for substantially all identical expressible nucleotide sequences – different promoters.

30

The common phenotype of a given sub-library may be at least one phenotype selected from the group comprising the ability to grow on unusual substrates, the ability to grow on sublethal concentration of toxins, the ability to grow at a high temperature, the ability to grow at a low temperature, the ability to grow at elevated

35      osmolality, the ability to grow at low osmolality, the ability to grow at high salinity, the

ability to grow at low salinity, the ability to grow at elevated metal concentrations, the ability to grow at high CO<sub>2</sub> concentrations, the ability to grow at low CO<sub>2</sub> concentrations, the ability to grow at high O<sub>2</sub> concentrations, the ability to grow at low O<sub>2</sub> concentrations, the ability to provide special spectral properties, the ability to provide a special colour, the ability to have a deviating GST activity, the ability to have a deviating P450 activity.

### **Size of library**

A library of cells may in principle comprise just two cells differing with respect to one of the features discussed below. However, normally a library comprises at least 20 individual cells, such as at least 50 individual cells. More preferably, a library comprises at least 100 individual cells, such as at least 1,000 cells, for example at least 10,000 cells such as at least 100,000 cells, for example at least 1,000,000 cells, such as at least 1,000,000,000 cells.

The number of cells in a sub-library depends on the selection criterion or criteria used. At the beginning a sub-library typically comprises less cells than a library, but the cells of the sub-library may be combined or allowed to sexually propagate to produce increased variation and in this way the number of different cells in a sub-library may increase.

### **Variation among cells**

The difference between cells in a library may be defined with reference to differences between expression cassettes, between concatemers or differences between promoters controlling the expression of an expressible nucleotide sequence.

Thus in a library according to the invention a concatemer of each cell may comprise at least a first cassette and a second cassette, said first cassette being different from said second cassette. More preferably substantially all cassettes of a concatemer in a given cell are different.

The difference between the expression cassettes, which may be reflected in the difference between concatemers in different cells may be a difference in the spacer sequences and/or the promoter, and/or the expressible nucleotide sequence and/or the intron and/or terminator sequence.

5

When the differences lie in the expressible nucleotide sequences these different expressible nucleotide sequences may come from the same or from different expression states. The different expression states may represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera. The different species are from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms. In this way cells and libraries representing an extremely wide array of gene combinations is obtained.

15

Preferably substantially all cells in a library are different. This increases the number of available combinations of expressible nucleotide sequences. Further variation may be obtained by having one library in cells of one mating type and another library in cells of another mating type. For yeast this may be obtained by having one library in  $\text{Mata}^-$  cells and another library in  $\text{Mata}^+$  cells. These may then be sexually crossed to obtain further variation.

20

According to an especially preferred embodiment of the invention the library comprises a random combination of promoter and expressible nucleotide sequences made from a two dimensional array of promoters and heterologous expressible nucleotide sequences. Thereby, it is possible to get – in principle – all expressible nucleotide sequences from a given pool represented in a library under the control of different promoters.

25

When each cell furthermore comprises an individual selection of combinations of promoters and heterologous expressible nucleotide sequences drawn individually from the same pool of promoters and heterologous expressible nucleotide sequences completely random combinations of promoter and expressible nucleotide sequences are inserted into all cells. Each expressible nucleotide sequence may

30



then be found in the library under the control of different promoters and in a number of combinations with a number of other expressible nucleotide sequences.

Each library may comprise at least 2 different independently controllable promoters, such as at least 3, for example at least 4, such as at least 5, for example at least 6, such as at least 7, for example at least 8, such as at least 9, for example at least 10, such as at least 15, for example at least 25, such as at least 50, for example at least 75, such as at least 100. The higher the number of promoters in the library, the number of sub-sets of genes may be constructed within any one cell and within any one library. Preferably the regulation of the promoters should not interact on each other. The absence of interaction sets an upper limit to the number of promoters that can be used under practical circumstances. However, new promoters are discovered and synthetic promoters are being developed continuously so it is likely that in the future combinations of different non-interacting promoters can be made.

At least one heterologous expressible nucleotide sequence may be found in at least 2 cells, such as at least 3 cells, for example at least 5 cells, such as at least 10 cells, for example at least 25 cell, such as at least 50 cells, for example at least 100 cells, such as at least 500 cells, for example at least 1000 cells. By having the same expressible nucleotide represented in several preferably in many cells, any one expressible nucleotide sequence may be found in many combinations with different expressible nucleotide sequences.

The combination of promoter and expressible nucleotide sequences in any one cell may be laid out so that at least one cell comprises a group of heterologous expressible nucleotide sequences under the control of a first promoter, the group comprising at least 5 heterologous expressible nucleotide sequences, such as at least 10 heterologous expressible nucleotide sequences, for example at least 15 heterologous expressible nucleotide sequences, such as at least 25 heterologous expressible nucleotide sequences, for example at least 50 heterologous expressible nucleotide sequences, such as at least 75 heterologous expressible nucleotide sequences, for example at least 100 heterologous expressible nucleotide sequences, such as at least 250 heterologous expressible nucleotide sequences, for example at least 500 heterologous expressible nucleotide sequences. Thereby a

sub-set of expressible nucleotide sequences of different size can be turned on and off in the cells.

5 By furthermore having in a cell at least a second group of heterologous expressible nucleotide sequences under the independent control of second promoter, such as at least a third group of heterologous expressible nucleotide sequences under the independent control of a third promoter, for example at least a fourth group of heterologous expressible nucleotide sequences under the independent control of a fourth promoter, such as at least a fifth group of heterologous expressible nucleotide sequences under the independent control of a fifth promoter, for example at least a sixth group of heterologous expressible nucleotide sequences under the independent control of a sixth promoter, such as at least a seventh group of heterologous expressible nucleotide sequences under the independent control of a seventh promoter, such as at least a eighth group of heterologous expressible nucleotide sequences under the independent control of a eighth promoter, for example at least a ninth group of heterologous expressible nucleotide sequences under the independent control of a ninth promoter, such as at least a tenth group of heterologous expressible nucleotide sequences under the independent control of a tenth promoter, groups of expressible nucleotide sequences, sub-sets, may be turned on and off in the cells.

#### **Origin of expressible nucleotide sequences**

The expressible nucleotide sequences that can be inserted into the vectors, concatemers, and cells according to this invention encompass any type of nucleotide such as RNA, DNA. Such a nucleotide sequence could be obtained e.g. from cDNA, which by its nature is expressible. But it is also possible to use sequences of genomic DNA, coding for specific genes. Preferably, the expressible nucleotide sequences correspond to full length genes such as substantially full length cDNA, but nucleotide sequences coding for shorter peptides than the original full length mRNAs may also be used. Shorter peptides may still retain the catalytic activity similar to that of the native proteins.

Another way to obtain expressible nucleotide sequences is through chemical synthesis of nucleotide sequences coding for known peptide or protein sequences. Thus the expressible DNA sequences does not have to be a naturally occurring

sequence, although it may be preferable for practical purposes to primarily use naturally occurring nucleotide sequences. Whether the DNA is single or double stranded will depend on the vector system used.

- 5 In most cases the orientation with respect to the promoter of an expressible nucleotide sequence will be such that the coding strand is transcribed into a proper mRNA. It is however conceivable that the sequence may be reversed generating an antisense transcript in order to block expression of a specific gene.

## 10 **Cassettes**

An important aspect of the invention concerns a cassette of nucleotides in a highly ordered sequence, the cassette having the general formula in 5'→3' direction:

[RS1-RS2-SP-PR-CS-TR-SP-RS2'-RS1']

- 15 wherein RS1 and RS1' denote restriction sites, RS2 and RS2' denote restriction sites different from RS1 and RS1', SP individually denotes a spacer sequence of at least two nucleotides, PR denotes a promoter, CS denotes a cloning site, and TR denotes a terminator.

- 20 It is an advantage to have two different restriction sites flanking both sides of the expression construct. By treating the primary vectors with restriction enzymes cleaving both restriction sites, the expression construct and the primary vector will be left with two non-compatible ends. This facilitates a concatenation process, since the empty vectors do not participate in the concatenation of expression constructs.

25

## **Restriction sites**

- In principle, any restriction site, for which a restriction enzyme is known can be used. These include the restriction enzymes generally known and used in the field of  
30 molecular biology such as those described in Sambrook, Fritsch, Maniatis, "A laboratory Manual", 2<sup>nd</sup> edition. Cold Spring Harbor Laboratory Press, 1989.

- The restriction site recognition sequences preferably are of a substantial length, so that the likelihood of occurrence of an identical restriction site within the cloned  
35 oligonucleotide is minimised. Thus the first restriction site may comprise at least 6

bases, but more preferably the recognition sequence comprises at least 7 or 8 bases. Restriction sites having 7 or more non N bases in the recognition sequence are generally known as "rare restriction sites" (see example 6). However, the recognition sequence may also be at least 10 bases, such as at least 15 bases, for example at least 16 bases, such as at least 17 bases, for example at least 18 bases, such as at least 18 bases, for example at least 19 bases, for example at least 20 bases, such as at least 21 bases, for example at least 22 bases, such as at least 23 bases, for example at least 25 bases, such as at least 30 bases, for example at least 35 bases, such as at least 40 bases, for example at least 45 bases, such as at least 50 bases.

Preferably the first restriction site RS1 and RS1' is recognised by a restriction enzyme generating blunt ends of the double stranded nucleotide sequences. By generating blunt ends at this site, the risk that the vector participates in a subsequent concatenation is greatly reduced. The first restriction site may also give rise to sticky ends, but these are then preferably non-compatible with the sticky ends resulting from the second restriction site, RS2 and RS2' and with the sticky ends in the AC.

According to a preferred embodiment of the invention, the second restriction site, RS2 and RS2' comprises a rare restriction site. Thus, the longer the recognition sequence of the rare restriction site the more rare it is and the less likely is it that the restriction enzyme recognising it will cleave the nucleotide sequence at other – undesired – positions.

The rare restriction site may furthermore serve as a PCR priming site. Thereby it is possible to copy the cassettes via PCR techniques and thus indirectly "excise" the cassettes from a vector.

### 30 **Spacer sequence**

The spacer sequence located between the RS2 and the PR sequence is preferably a non-transcribed spacer sequence. The purpose of the spacer sequence(s) is to minimise recombination between different concatemers present in the same cell or between cassettes present in the same concatemer, but it may also serve the pur-

pose of making the nucleotide sequences in the cassettes more "host" like. A further purpose of the spacer sequence is to reduce the occurrence of hairpin formation between adjacent palindromic sequences, which may occur when cassettes are assembled head to head or tail to tail. Spacer sequences may also be convenient for introducing short conserved nucleotide sequences that may serve e.g. as PCR primer sites or as target for hybridization to e.g. nucleic acid or PNA or LNA probes allowing affinity purification of cassettes.

The cassette may also optionally comprise another spacer sequence of at least two nucleotides between TR and RS2. When cassettes are cut out from a vector and concatenated into concatemers of cassettes, the spacer sequences together ensure that there is a certain distance between two successive identical promoter and/or terminator sequences. This distance may comprise at least 50 bases, such as at least 60 bases, for example at least 75 bases, such as at least 100 bases, for example at least 150 bases, such as at least 200 bases, for example at least 250 bases, such as at least 300 bases, for example at least 400 bases, for example at least 500 bases, such as at least 750 bases, for example at least 1000 bases, such as at least 1100 bases, for example at least 1200 bases, such as at least 1300 bases, for example at least 1400 bases, such as at least 1500 bases, for example at least 1600 bases, such as at least 1700 bases, for example at least 1800 bases, such as at least 1900 bases, for example at least 2000 bases, such as at least 2100 bases, for example at least 2200 bases, such as at least 2300 bases, for example at least 2400 bases, such as at least 2500 bases, for example at least 2600 bases, such as at least 2700 bases, for example at least 2800 bases, such as at least 2900 bases, for example at least 3000 bases, such as at least 3200 bases, for example at least 3500 bases, such as at least 3800 bases, for example at least 4000 bases, such as at least 4500 bases, for example at least 5000 bases, such as at least 6000 bases.

The number of the nucleotides between the spacer located 5' to the PR sequence and the one located 3' to the TR sequence may be any. However, it may be advantageous to ensure that at least one of the spacer sequences comprises between 100 and 2500 bases, preferably between 200 and 2300 bases, more preferably between 300 and 2100 bases, such as between 400 and 1900 bases, more preferably between 500 and 1700 bases, such as between 600 and 1500 bases, more preferably between 700 and 1400 bases.

If the intended host cell is yeast, the spacers present in a concatemer should preferably comprise a combination of a few ARSes with varying lambda phage DNA fragments.

5

Preferred examples of spacer sequences include but are not limited to: Lambda phage DNA, prokaryotic genomic DNA such as *E. coli* genomic DNA, ARSes.

### Promoter

10 A promoter is a DNA sequence to which RNA polymerase binds and initiates transcription. The promoter determines the polarity of the transcript by specifying which strand will be transcribed.

15

- Bacterial promoters normally consist of -35 and -10 (relative to the transcriptional start) consensus sequences which are bound by a specific sigma factor and RNA polymerase.

20

- Eukaryotic promoters are more complex. Most promoters utilized in expression vectors are transcribed by RNA polymerase II. General transcription factors (GTFs) first bind specific sequences near the transcriptional start and then recruit the binding of RNA polymerase II. In addition to these minimal promoter elements, small sequence elements are recognized specifically by modular DNA-binding / trans-activating proteins (e.g. AP-1, SP-1) which regulate the activity of a given promoter.

25

- Viral promoters may serve the same function as bacterial and eukaryotic promoters. Upon viral infection of their host, viral promoters direct transcription either by using host transcriptional machinery or by supplying virally encoded enzymes to substitute part of the host machinery. Viral promoters are recognised by the transcriptional machinery of a large number of host organisms and are therefore often used in cloning and expression vectors.

30 Promoters may furthermore comprise regulatory elements, which are DNA sequence elements which act in conjunction with promoters and bind either repressors (e.g., lacO/ LAC Iq repressor system in *E. coli*) or inducers (e.g., gal1 /GAL4 inducer system in yeast). In either case, transcription is virtually "shut off" until the promoter is derepressed or induced, at which point transcription is "turned-

on". The choice of promoter in the cassette is primarily dependent on the host organism into which the cassette is intended to be inserted. An important requirement to this end is that the promoter should preferably be capable of functioning in the host cell, in which the expressible nucleotide sequence is to be expressed.

Preferably the promoter is an externally controllable promoter, such as an inducible promoter and/or a repressible promoter. The promoter may be either controllable (repressible/inducible) by chemicals such as the absence/presence of chemical inducers, e.g. metabolites, substrates, metals, hormones, sugars. The promoter may likewise be controllable by certain physical parameters such as temperature, pH, redox status, growth stage, developmental stage, or the promoter may be inducible/repressible by a synthetic inducer/repressor such as the gal inducer.

In order to avoid unintentional interference with the gene regulation systems of the host cell, and in order to improve controllability of the co-ordinated gene expression the promoter is preferably a synthetic promoter. Suitable promoters are described in US 5,798,227, US 5,667,986. Principles for designing suitable synthetic eukaryotic promoters are disclosed in US 5,559,027, US 5,877,018 or US 6,072,050.

Synthetic inducible eukaryotic promoters for the regulation of transcription of a gene may achieve improved levels of protein expression and lower basal levels of gene expression. Such promoters preferably contain at least two different classes of regulatory elements, usually by modification of a native promoter containing one of the inducible elements by inserting the other of the inducible elements. For example, additional metal responsive elements (MREs) and/or glucocorticoid responsive elements (GREs) may be provided to native promoters. Additionally, one or more constitutive elements may be functionally disabled to provide the lower basal levels of gene expression.

Preferred examples of promoters include but is not limited to those promoters being induced and/or repressed by any factor selected from the group comprising carbohydrates, e.g. galactose; low inorganic phosphate levels; temperature, e.g. low or high temperature shift; metals or metal ions, e.g. copper ions; hormones, e.g. dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-

status; growth stage, e.g. developmental stage; synthetic inducers, e.g. gal inducer. Examples of such promoters include ADH 1, PGK 1, GAP 491, TPI, PYK, ENO, PMA 1, PHO5, GAL 1, GAL 2, GAL 10, MET25, ADH2, MEL 1, CUP 1, HSE, AOX, MOX, SV40, CaMV, Opaque-2, GRE, ARE, PGK/ARE hybrid, CYC/GRE hybrid, TPI/ $\alpha$ 2 operator, AOX 1, MOX A.

More preferably, however the promoter is selected from hybrid promoters such as PGK/ARE hybrid, CYC/GRE hybrid or from synthetic promoters. Such promoters can be controlled without interfering too much with the regulation of native genes in the expression host.

### **Yeast promoters**

In the following, examples of known yeast promoters that may be used in conjunction with the present invention are shown. The examples are by no way limiting and only serve to indicate to the skilled practitioner how to select or design promoters that are useful according to the present invention.

Although numerous transcriptional promoters which are functional in yeasts have been described in the literature, only some of them have proved effective for the production of polypeptides by the recombinant route. There may be mentioned in particular the promoters of the PGK genes (3-phosphoglycerate kinase, TDH genes encoding GAPDH (Glyceraldehyde phosphate dehydrogenase), TEF1 genes (Elongation factor 1), MF $\alpha$ 1 ( $\alpha$  sex pheromone precursor) which are considered as strong constitutive promoters or alternatively the regulatable promoter CYC1 which is repressed in the presence of glucose or PHO5 which can be regulated by thiamine. However, for reasons which are often unexplained, they do not always allow the effective expression of the genes which they control. In this context, it is always advantageous to be able to have new promoters in order to generate new effective host/vector systems. Furthermore, having a choice of effective promoters in a given cell also makes it possible to envisage the production of multiple proteins in this same cell (for example several enzymes of the same metabolic chain) while avoiding the problems of recombination between homologous sequences.



In general, a promoter region is situated in the 5' region of the genes and comprises all the elements allowing the transcription of a DNA fragment placed under their control, in particular:

- 5 (1) a so-called minimal promoter region comprising the TATA box and the site of initiation of transcription, which determines the position of the site of initiation as well as the basal level of transcription. In *Saccharomyces cerevisiae*, the length of the minimal promoter region is relatively variable. Indeed, the exact location of the TATA box varies from one gene to another and may be situated from -40 to -  
10 120 nucleotides upstream of the site of the initiation (Chen and Struhl, 1985, EMBO J., 4, 3273-3280)
- (2) sequences situated upstream of the TATA box (immediately upstream up to  
15 several hundreds of nucleotides) which make it possible to ensure an effective level of transcription either constitutively (relatively constant level of transcription all along the cell cycle, regardless of the conditions of culture) or in a regulatable manner (activation of transcription in the presence of an activator and/or  
20 repression in the presence of a repressor). These sequences, may be of several types: activator, inhibitor, enhancer, inducer, repressor and may respond to cellular factors or varied culture conditions.

20

Examples of such promoters are the ZZA1 and ZZA2 promoters disclosed in US 5,641,661, the EF1- $\alpha$  protein promoter and the ribosomal protein S7 gene promoter disclosed in WO 97/44470,, the COX 4 promoter and two unknown promoters (SEQ ID No: 1 and 2 in the document) disclosed in US 5,952,195. Other useful promoters  
25 include the HSP150 promoter disclosed in WO 98/54339 and the SV40 and RSV promoters disclosed in US 4,870,013 as well as the PyK and GAPDH promoters disclosed in EP 0 329 203 A1.

### **Synthetic yeast promoters**

- 30 More preferably the invention employs the use of synthetic promoters. Synthetic promoters are often constructed by combining the minimal promoter region of one gene with the upstream regulating sequences of another gene. Enhanced promoter control may be obtained by modifying specific sequences in the upstream regulating sequences, e.g. through substitution or deletion or through inserting multiple copies  
35 of specific regulating sequences. One advantage of using synthetic promoters is that

they may be controlled without interfering too much with the native promoters of the host cell.

5 One such synthetic yeast promoter comprises promoters or promoter elements of two different yeast-derived genes, yeast killer toxin leader peptide, and amino terminus of IL-1 $\beta$  (WO 98/54339).

10 Another example of a yeast synthetic promoter is disclosed in US 5,436,136 (Hinnen et al), which concerns a yeast hybrid promoter including a 5' upstream promoter element comprising upstream activation site(s) of the yeast PHO5 gene and a 3' downstream promoter element of the yeast GAPDH gene starting at nucleotide -300 to -180 and ending at nucleotide -1 of the GAPDH gene.

15 Another example of a yeast synthetic promoter is disclosed in US 5,089,398 (Rosenberg et al). This disclosure describes a promoter with the general formula - (P.R.(2)-P.R.(1))-

wherein:

20 P.R.(1) is the promoter region proximal to the coding sequence and having the transcription initiation site, the RNA polymerase binding site, and including the TATA box, the CAAT sequence, as well as translational regulatory signals, e.g., capping sequence, as appropriate;

P.R.(2) is the promoter region joined to the 5'-end of P.R.(1) associated with enhancing the efficiency of transcription of the RNA polymerase binding region;

25 In US 4,945,046 (Horii et al) discloses a further example of how to design a synthetic yeast promoter. This specific promoter comprises promoter elements derived both from yeast and from a mammal. The hybrid promoter consists essentially of *Saccharomyces cerevisiae* PHO5 or GAP-DH promoter from which the upstream activation site (UAS) has been deleted and replaced by the early enhancer region derived from SV40 virus.

30

### **Cloning site**

The cloning site in the cassette in the primary vector should be designed so that any nucleotide sequence can be cloned into it.

35

The cloning site in the cassette preferably allows directional cloning. Hereby is ensured that transcription in a host cell is performed from the coding strand in the intended direction and that the translated peptide is identical to the peptide for which the original nucleotide sequence codes.

5

However according to some embodiments it may be advantageous to insert the sequence in opposite direction. According to these embodiments, so-called antisense constructs may be inserted which prevent functional expression of specific genes involved in specific pathways. Thereby it may become possible to divert metabolic intermediates from a prevalent pathway to another less dominant pathway.

10

The cloning site in the cassette may comprise multiple cloning sites, generally known as MCS or polylinker sites, which is a synthetic DNA sequence encoding a series of restriction endonuclease recognition sites. These sites are engineered for convenient cloning of DNA into a vector at a specific position and for directional cloning of the insert.

15

Cloning of cDNA does not have to involve the use of restriction enzymes. Other alternative systems include but are not limited to:

20

- Creator™ Cre-loxP system from Clontech, which uses recombination and loxP sites
- use of Lambda attachment sites (att-λ), such as the Gateway™ system from Life Technologies.

25

Both of these systems are directional.

### **Terminator**

The role of the terminator sequence is to limit transcription to the length of the coding sequence. An optimal terminator sequence is thus one, which is capable of performing this act in the host cell.

30

In prokaryotes, sequences known as transcriptional terminators signal the RNA polymerase to release the DNA template and stop transcription of the nascent RNA.

In eukaryotes, RNA molecules are transcribed well beyond the end of the mature mRNA molecule. New transcripts are enzymatically cleaved and modified by the addition of a long sequence of adenylic acid residues known as the poly-A tail. A polyadenylation consensus sequence is located about 10 to 30 bases upstream from the actual cleavage site.

Preferred examples of yeast derived terminator sequences include, but are not limited to: ADN1, CYC1, GPD, ADH1 alcohol dehydrogenase.

### **Intron**

Optionally, the cassette in the vector comprises an intron sequence, which may be located 5' or 3' to the expressible nucleotide sequence. The design and layout of introns is well known in the art. The choice of intron design largely depends on the intended host cell, in which the expressible nucleotide sequence is eventually to be expressed. The effects of having intron sequence in the expression cassettes are those generally associated with intron sequences.

Examples of yeast introns can be found in the literature and in specific databases such as Ares Lab Yeast Intron Database (Version 2.1) as updated on 15 April 2000. Earlier versions of the database as well as extracts of the database have been published in: "Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*." by Spingola M, Grate L, Haussler D, Ares M Jr. (RNA 1999 Feb;5(2):221-34) and "Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast." by Davis CA, Grate L, Spingola M, Ares M Jr, (Nucleic Acids Res 2000 Apr 15;28(8):1700-6).

### **Primary vectors (entry vectors)**

By the term entry vector is meant a vector for storing and amplifying cDNA or other expressible nucleotide sequences using the cassettes according to the present invention. The primary vectors are preferably able to propagate in *E. coli* or any other suitable standard host cell. It should preferably be amplifiable and amenable to standard normalisation and enrichment procedures.

The primary vector may be of any type of DNA that has the basic requirements of a) being able to replicate itself in at least one suitable host organism and b) allows insertion of foreign DNA which is then replicated together with the vector and c) preferably allows selection of vector molecules that contain insertions of said foreign DNA. In a preferred embodiment the vector is able to replicate in standard hosts like yeasts, and bacteria and it should preferably have a high copy number per host cell. It is also preferred that the vector in addition to a host specific origin of replication, contains an origin of replication for a single stranded virus, such as e.g. the f1 origin for filamentous phages. This will allow the production of single stranded nucleic acid sequences. A vast number of cloning vectors have been described which are commonly used and references may be given to e.g. Sambrook, J; Fritsch, E.F; and Maniatis T. (1989) Molecular Cloning: A laboratory manual. Cold Spring Harbour Laboratory Press, USA, Netherlands Culture Collection of Bacteria ([www.cbs.knaw.nl/NCCB/collection.htm](http://www.cbs.knaw.nl/NCCB/collection.htm)) or Department of Microbial Genetics, National Institute of Genetics, Yata 1111 Mishima Shizuoka 411-8540, Japan ([www.shigen.nig.ac.jp/cvector/cvector.html](http://www.shigen.nig.ac.jp/cvector/cvector.html)). A few type-examples that are the parents of many popular derivatives are M13mp10, pUC18, Lambda gt 10, and pYAC4. Examples of primary vectors include but are not limited to M13K07, pBR322, pUC18, pUC19, pUC118, pUC119, pSP64, pSP65, pGEM-3, pGEM-3Z, pGEM-3Zf(-), pGEM-4, pGEM-4Z,  $\pi$ AN13, pBluescript II, CHARON 4A,  $\lambda^+$ , CHARON 21A, CHARON 32, CHARON 33, CHARON 34, CHARON 35, CHARON 40, EMBL3A,  $\lambda$ 2001,  $\lambda$ DASH,  $\lambda$ FIX,  $\lambda$ gt10,  $\lambda$ gt11,  $\lambda$ gt18,  $\lambda$ gt20,  $\lambda$ gt22,  $\lambda$ ORF8,  $\lambda$ ZAP/R, pJB8, c2RB, pcos1EMBL

Methods for cloning of cDNA or genomic DNA into a vector are well known in the art. Reference may be given to J. Sambrook, E.F. Fritsch, T. Maniatis: Molecular Cloning, A Laboratory Manual (2<sup>nd</sup> edition, Cold Spring Harbor Laboratory Press, 1989).

One example of a circular model entry vector is described in Figure 3. The vector, EVE contains the expression cassette, R1-R2-Spacer-Promoter-Multi Cloning Site-Terminator-Spacer-R2-R1. The vector furthermore contains a gene for ampicillin resistance, AmpR, and an origin of replication for E.coli, ColE1.

The entry vectors EVE4, EVE5, and EVE8 shown in Figures 4, 5, and 6. These all contain SrfI as R1 and AscI as R2. Both of these sites are palindromic and are regarded as rare restriction sites having 8 bases in the recognition sequence. The vectors furthermore contain the AmpR ampicillin resistance gene, and the ColE1 origin or replication for E.coli as well as f1, which is an origin of replication for filamentous phages, such as M13. EVE4 (Fig. 4) contains the MET25 promoter and the ADH1 terminator. Spacer 1 and spacer 2 are short sequences deriving from the multiple cloning site, MCS. EVE5 (Fig. 5) contains the CUP1 promoter and the ADH1 terminator. EVE8 (Fig. 6) contains the CUP1 promoter and the ADH1 terminator. The spacers of EVE8 are a 550 bp lambda phage DNA (spacer 3) and an ARS sequence from yeast (spacer 4).

#### **Nucleotide library (entry library)**

Methods as well as suitable vectors and host cells for constructing and maintaining a library of nucleotide sequences in a cell are well known in the art. The primary requirement for the library is that it should be possible to store and amplify in it a number of primary vectors (constructs) according to this invention, the vectors (constructs) comprising expressible nucleotide sequences from at least one expression state and wherein at least two vectors (constructs) are different.

One specific example of such a library is the well known and widely employed cDNA libraries. The advantage of the cDNA library is mainly that it contains only DNA sequences corresponding to transcribed messenger RNA in a cell. Suitable methods are also present to purify the isolated mRNA or the synthesised cDNA so that only substantially full-length cDNA is cloned into the library.

Methods for optimisation of the process to yield substantially full length cDNA may comprise size selection, e.g. electrophoresis, chromatography, precipitation or may comprise ways of increasing the likelihood of getting full length cDNAs, e.g. the SMART™ method (Clontech) or the CapTrap™ method (Stratagene).

Preferably the method for making the nucleotide library comprises obtaining a substantially full length cDNA population comprising a normalised representation of cDNA species. More preferably a substantially full length cDNA population

comprises a normalised representation of cDNA species characteristic of a given expression state.

5 Normalisation reduces the redundancy of clones representing abundant mRNA species and increases the relative representation of clones from rare mRNA species.

10 Methods for normalisation of cDNA libraries are well known in the art. Reference may be given to suitable protocols for normalisation such as those described in US 5,763,239 (DIVERSA) and WO 95/08647 and WO 95/11986. and Bonaldo, Lennon, Soares, Genome Research 1996, 6:791-806; Ali, Holloway, Taylor, Plant Mol Biol Reporter, 2000, 18:123-132.

15 Enrichment methods are used to isolate clones representing mRNA which are characteristic of a particular expression state. A number of variations of the method broadly termed as subtractive hybridisation are known in the art. Reference may be given to Sive, John, Nucleic Acid Res, 1988, 16:10937; Diatchenko, Lau, Campbell et al, PNAS, 1996, 93:6025-6030; Carninci, Shibata, Hayatsu, Genome Res, 2000, 10:1617-30, Bonaldo, Lennon, Soares, Genome Research 1996, 6:791-806; Ali, 20 Holloway, Taylor, Plant Mol Biol Reporter, 2000, 18:123-132. For example, enrichment may be achieved by doing additional rounds of hybridization similar to normalization procedures, using e.g. cDNA from a library of abundant clones or simply a library representing the uninduced state as a driver against a tester library from the induced state. Alternatively mRNA or PCR amplified cDNA derived from the 25 expression state of choice can be used to subtract common sequences from a tester library. The choice of driver and tester population will depend on the nature of target expressible nucleotide sequences in each particular experiment

30 In the library an expressible nucleotide sequence coding for one peptide is preferably found in different but similar vectors under the control of different promoters. Preferably the library comprises at least three primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of three different promoters. More preferably the library comprises at least four primary vectors with an expressible nucleotide sequence coding for the same peptide under 35 the control of four different promoters. More preferably the library comprises at least

five primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of five different promoters, such as comprises at least six primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of six different promoters, for example comprises at least seven primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of seven different promoters, for example comprises at least eight primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of eight different promoters, such as comprises at least nine primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of nine different promoters, for example comprises at least ten primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of ten different promoters.

The expressible nucleotide sequence coding for the same peptide preferably comprises essentially the same nucleotide sequence, more preferably the same nucleotide sequence.

By having a library with what may be termed one gene under the control of a number of different promoters in different vectors, it is possible to construct from the nucleotide library an array of combinations of genes and promoters. Preferably, one library comprises a complete or substantially complete combination such as a two dimensional array of genes and promoters, wherein substantially all genes are found under the control of substantially all of a selected number of promoters.

According to another embodiment of the invention the nucleotide library comprises combinations of expressible nucleotide sequences combined in different vectors with different spacer sequences and/or different intron sequences. Thus any one expressible nucleotide sequence may be combined in a two, three, four or five dimensional array with different promoters and/or different spacers and/or different introns and/or different terminators. The two, three, four or five dimensional array may be complete or incomplete, since not all combinations will have to be present.

The library may suitably be maintained in a host cell comprising prokaryotic cells or eukaryotic cells. Preferred prokaryotic host organisms may include but are not



limited to *Escherichia coli*, *Bacillus subtilis*, *Streptomyces lividans*, *Streptomyces coelicolor*, *Pseudomonas aeruginosa*, *Myxococcus xanthus*.

5 Yeast species such as *Saccharomyces cerevisiae* (budding yeast),  
Schizosaccharomyces pombe (fission yeast), *Pichia pastoris*, and *Hansenula polymorpha* (methylotropic yeasts) may also be used. Filamentous ascomycetes, such as *Neurospora crassa* and *Aspergillus nidulans* may also be used. Plant cells such as those derived from *Nicotiana* and *Arabidopsis* are preferred. Preferred mammalian host cells include but are not limited to those derived from humans,  
10 monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa etc (see Kriegler M. in "Gene Transfer and Expression: A Laboratory Manual", New York, Freeman & Co. 1990).

### Concatemers

15 A concatemer is a series of linked units. In the present context a concatemer is used to denote a number of serially linked nucleotide cassettes, wherein at least two of the serially linked nucleotide units comprises a cassette having the basic structure

[rs<sub>2</sub>-SP-PR-X-TR-SP-rs<sub>1</sub>]

20 wherein

rs<sub>1</sub> and rs<sub>2</sub> together denote a restriction site,

SP individually denotes a spacer of at least two nucleotide bases,

PR denotes a promoter, capable of functioning in a cell,

X denotes an expressible nucleotide sequence,

25 TR denotes a terminator, and

SP individually denotes a spacer of at least two nucleotide bases.

Optionally the cassettes comprise an intron sequence between the promoter and the expressible nucleotide sequence and/or between the terminator and the expressible  
30 sequence.

The expressible nucleotide sequence in the cassettes of the concatemer may comprise a DNA sequence selected from the group comprising cDNA and genomic DNA.

35

According to one aspect of the invention, a concatemer comprises cassettes with expressible nucleotide from different expression states, so that non-naturally occurring combinations or non-native combinations of expressible nucleotide sequences are obtained. These different expression states may represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera. The different species may be from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.

For example, the expressible nucleotide sequences may originate from eukaryotes such as mammals such as humans, mice or whale, from reptiles such as snakes crocodiles or turtles, from tunicates such as sea squirts, from lepidoptera such as butterflies and moths, from coelenterates such as jellyfish, anenomes, or corals, from fish such as bony and cartilaginous fish, from plants such as dicots, e.g. coffee, oak or monocots such as grasses, lilies, and orchids; from lower plants such as algae and gingko, from higher fungi such as terrestrial fruiting fungi, from marine actinomycetes. The expressible nucleotide sequences may also originate from protozoans such as malaria or trypanosomes, or from prokaryotes such as E. coli or archaeobacteria. Furthermore, the expressible nucleotide sequences may originate from one or more preferably from more expression states from the species and genera listed in the table below.

Bacteria	Streptomyces, Micromonospora, Norcadia, Actinomadura, Actinoplanes, Streptosporangium, Microbispora, Kitasatosporium, Azobacterium, Rhizobium, Achromobacterium, Enterobacterium, Brucella, Micrococcus, Lactobacillus, Bacillus (B.t. toxins), Clostridium (toxins), Brevibacterium, Pseudomonas, Aerobacter, Vibrio, Halobacterium, Mycoplasma, Cytophaga, Myxococcus
Fungi	Amanita muscaria (fly agaric, ibotenic acid, muscimol), Psilocybe (psilocybin) Physarium, Fuligo, Mucor, Phytophthora, Rhizopus, Aspergillus, Penicillium (penicillin), Coprinus, Phanerochaete, Acremonium (Cephalosporin), Trochoderma, Helminthosporium, Fusarium, Alternaria, Myrothecium, Saccharomyces
Algae	Digenea simplex (kainic acid, antihelminthic), Laminaria anqustata (laminine, hypotensive)

	Lichens	Usnea fasciata (vulpinicacid, antimicrobial; usnic acid, antitumor)
5	Higher Plants	Artemisia (artemisinin), Coleus (forskolin), Desmodium (K channel agonist), Catharanthus (Vinca alkaloids), Digitalis (cardiac glycosides), Podophyllum (podophyllotoxin), Taxus (taxol), Cephalotaxus (homoharringtonine), Camptotheca (Camptothecin), Camellia sinensis (Tea), Cannabis indica, Cannabis sativa (Hemp), Erythroxylum coca (Coca), Lophophora williamsii (Peyote), Myristica fragrans (Nutmeg), Nicotiana, Papaver somniferum (Opium Poppy), Phalaris arundinacea (Reed canary grass)
10	Protozoa	Ptychodiscus brevis; Dinoflagellates (brevitoxin, cardiovascular)
15	Sponges	Microciona prolifera (ectyonin, antimicrobial) Cryptotethya crypta (D-arabino furanosides)
	Coelenterata	Portuguese Man o War & other jellyfish and medusoid toxins.
20	Corals	Pseudoterogonia species (Pseudoteracins, anti-inflammatory), Erythropodium (erythrolides, anti-inflammatory)
	Aschelminths	Nematode secretory compounds
	Molluscs	Conus toxins, sea slug toxins, cephalopod neurotransmitters, squid inks
25	Annelida	Lumbriconereis heteropa (nereistoxin, insecticidal)
	Arachnids	Dolomedes ("fishing spider" venoms)
30	Crustacea	Xenobalanus (skin adhesives)
	Insects	Epilachna (mexican bean beetle alkaloids)
	Spinunculida	Bonellia viridis (bonellin, neuroactive)
35	Bryozoans	Bugula neritina (bryostatins, anti cancer)
	Echinoderms	Crinoid chemistry
40	Tunicates	Trididemnum solidum (didemnin, anti-tumor and anti-viral; Ecteinascidia turbinata ecteinascidins, anti-tumor)
	Vertebrates	Eptatretus stoutii (eptatretin, cardioactive), Trachinus draco (proteinaceous toxins, reduce blood pressure, respiration and reduce heart rate). Dendrobatid frogs

(batrachotoxins, pumiliotoxins, histrionicotoxins, and other polyamines); Snake venom toxins; *Ornithorhynchus anatinus* (duck-billed platypus venom), modified carotenoids, retinoids and steroids; Avians: histrionicotoxins, modified carotenoids, retinoids and steroids

5

According to a preferred embodiment of the invention the concatemer comprises at least a first cassette and a second cassette, said first cassette being different from said second cassette. More preferably, the concatemer comprises cassettes, wherein substantially all cassettes are different. The difference between the cassettes may arise from differences between promoters, and/or expressible

10

nucleotide sequences, and/or spacers, and/or terminators, and/or introns.

The number of cassettes in a single concatemer is largely determined by the host species into which the concatemer is eventually to be inserted and the vector through which the insertion is carried out. The concatemer thus may comprise at least 10 cassettes, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least

15

20

2000 cassettes.

Each of the cassettes may be laid out as described above.

Once the concatemer has been assembled or concatenated it may be ligated into a suitable vector. Such a vector may advantageously comprise an artificial chromosome. The basic requirements for a functional artificial chromosome have been described in US 4,464,472, the contents of which is hereby incorporated by reference. An artificial chromosome or a functional minichromosome, as it may also be termed must comprise a DNA sequence capable of replication and stable mitotic maintenance in a host cell comprising a DNA segment coding for centromere-like activity during mitosis of said host and a DNA sequence coding for a replication site recognized by said host.

25

30

Suitable artificial chromosomes include a Yeast Artificial Chromosome (YAC) (see e.g. Murray et al, Nature 305:189-193; or US 4,464,472), a mega Yeast Artificial Chromosome (mega YAC), a Bacterial Artificial Chromosome (BAC), a mouse

35

artificial chromosome, a Mammalian Artificial Chromosome (MAC) (see e.g. US 6,133,503 or US 6,077,697), an Insect Artificial Chromosome (BUGAC), an Avian Artificial Chromosome (AVAC), a Bacteriophage Artificial Chromosome, a Baculovirus Artificial Chromosome, a plant artificial chromosome (US 5,270,201), a  
5 BIBAC vector (US 5,977,439) or a Human Artificial Chromosome (HAC).

The artificial chromosome is preferably so large that the host cell perceives it as a "real" chromosome and maintains it and transmits it as a chromosome. For yeast and other suitable host species, this will often correspond approximately to the size  
10 of the smallest native chromosome in the species. For *Saccharomyces*, the smallest chromosome has a size of 225 Kb.

MACs may be used to construct artificial chromosomes from other species, such as insect and fish species. The artificial chromosomes preferably are fully functional  
15 stable chromosomes. Two types of artificial chromosomes may be used. One type, referred to as SATACs [satellite artificial chromosomes] are stable heterochromatic chromosomes, and the other type are minichromosomes based on amplification of euchromatin.

20 Mammalian artificial chromosomes provide extra-genomic specific integration sites for introduction of genes encoding proteins of interest and permit megabase size DNA integration, such as integration of concatemers according to the invention.

According to another embodiment of the invention, the concatemer may be  
25 integrated into the host chromosomes or cloned into other types of vectors, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.

A preferable artificial chromosome vector is one that is capable of being conditionally amplified in the host cell, e.g. in yeast. The amplification preferably is at  
30 least a 10 fold amplification. Furthermore, it is advantageous that the cloning site of the artificial chromosome vector can be modified to comprise the same restriction site as the one bordering the cassettes described above, i.e. RS2 and/or RS2'.

## Concatenation

Cassettes to be concatenated are normally excised from a vector either by digestion with restriction enzymes or by PCR. After excision the cassettes may be separated from the vector through size fractionation such as gel filtration or through tagging of known sequences in the cassettes. The isolated cassettes may then be joined together either through interaction between sticky ends or through ligation of blunt ends.

Single-stranded compatible ends may be created by digestion with restriction enzymes. For concatenation a preferred enzyme for excising the cassettes would be a rare cutter, i.e. an enzyme that recognises a sequence of 7 or more nucleotides. Examples of enzymes that cut very rarely are the meganucleases, many of which are intron encoded, like e.g. I-Ceu I, I-Sce I, I-Ppo I, and PI-Psp I (see example 6d for more). Other preferred enzymes recognize a sequence of 8 nucleotides like e.g. Asc I, AsiS I, CciN I, CspB I, Fse I, MchA I, Not I, Pac I, Sbf I, Sda I, Sgf I, SgrA I, Sse232 I, and Sse8387 I, all of which create single stranded, palindromic compatible ends.

Other preferred rare cutters, which may also be used to control orientation of individual cassettes in the concatemer are enzymes that recognize non-palindromic sequences like e.g. Aar I, Sap I, Sfi I, Sdi I, and Vpa (see example 6c for more).

Alternatively, cassettes can be prepared by the addition of restriction sites to the ends, e.g. by PCR or ligation to linkers (short synthetic dsDNA molecules). Restriction enzymes are continuously being isolated and characterised and it is anticipated that many of such novel enzymes can be used to generate single-stranded compatible ends according to the present invention.

It is conceivable that single stranded compatible ends can be made by cleaving the vector with synthetic cutters. Thus, a reactive chemical group that will normally be able to cleave DNA unspecifically can cut at specific positions when coupled to another molecule that recognises and binds to specific sequences. Examples of molecules that recognise specific dsDNA sequences are DNA, PNA, LNA, phosphothioates, peptides, and amides. See e.g. Armitage, B.(1998) Chem. Rev.

98: 1171-1200, who describes photocleavage using e.g. anthraquinone and UV light; Dervan P.B. & Bürli R.W. (1999) Curr. Opin. Chem. Biol. 3: 688-93 describes the specific binding of polyamides to DNA; Nielsen, P.E. (2001) Curr. Opin. Biotechnol. 12: 16-20 describes the specific binding of PNA to DNA, and Chemical Reviews special thematic issue: RNA/DNA Cleavage (1998) vol. 98 (3) Bashkin J.K. (ed.) ACS publications, describes several examples of chemical DNA cleavers.

Single-stranded compatible ends may also be created by using e.g. PCR primers including dUTP and then treating the PCR product with Uracil-DNA glycosylase (Ref: US 5,035,996) to degrade part of the primer. Alternatively, compatible ends can be created by tailing both the vector and insert with complimentary nucleotides using Terminal Transferase (Chang, LMS, Bollum TJ (1971) J Biol Chem 246:909).

It is also conceivable that recombination can be used to generate concatemers, e.g. through the modification of techniques like the Creator™ system (Clontech) which uses the Cre-loxP mechanism (Sauer B 1993 Methods Enzymol 225:890-900) to directionally join DNA molecules by recombination or like the Gateway™ system (Life Technologies, US 5,888,732) using lambda *att* attachment sites for directional recombination (Landy A 1989, Ann Rev Biochem 58:913). It is envisaged that also lambda *cos* site dependent systems can be developed to allow concatenation.

More preferably the cassettes may be concatenated without an intervening purification step through excision from a vector with two restriction enzymes, one leaving sticky ends on the cassettes and the other one leaving blunt ends in the vectors. This is the preferred method for concatenation of cassettes from vectors having the basic structure of [RS1-RS2-SP-PR-X-TR-SP-RS2'-RS1'].

An alternative way of producing concatemers free of vector sequences would be to PCR amplify the cassettes from a single stranded primary vector. The PCR product must include the restriction sites RS2 and RS2' which are subsequently cleaved by its cognate enzyme(s). Concatenation can then be performed using the digested PCR product, essentially without interference from the single stranded primary vector template or the small double stranded fragments, which have been cut from the ends.

The concatemer may be assembled or concatenated by concatenation of at least two cassettes of nucleotide sequences each cassette comprising a first sticky end, a spacer sequence, a promoter, an expressible nucleotide sequence, a terminator, a spacer sequence, and a second sticky end. A flow chart of the procedure is shown in figure 2a.

Preferably concatenation further comprises

starting from a primary vector [RS1-RS2-SP-PR-X-TR-SP-RS2'-RS1'],

wherein X denotes an expressible nucleotide sequence,

RS1 and RS1' denote restriction sites,

RS2 and RS2' denote restriction sites different from RS1 and RS1',

SP individually denotes a spacer sequence of at least two nucleotides,

PR denotes a promoter,

TR denotes a terminator,

- i) cutting the primary vector with the aid of at least one restriction enzyme specific for RS2 and RS2' obtaining cassettes having the general formula [rs<sub>2</sub>-SP-PR-X-TR-SP-rs<sub>1</sub>] wherein rs<sub>1</sub> and rs<sub>2</sub> together denote a functional restriction site RS2 or RS2',
- ii) assembling the cut out cassettes through interaction between rs<sub>1</sub> and rs<sub>2</sub>.

In this way at least 10 cassettes can be concatenated, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000.

According to an especially preferred embodiment, vector arms each having a RS2 or RS2' in one end and a non-complementary overhang or a blunt end in the other end are added to the concatenation mixture together with the cassettes described above to further simplify the procedure (see Fig. 2b). One example of a suitable vector for providing vector arms is disclosed in Fig. 7 TRP1, URA3, and HIS3 are auxotrophic marker genes, and AmpR is an E. coli antibiotic marker gene. CEN4 is a centromer and TEL are telomeres. ARS1 and PMB1 allow replication in yeast and E. coli respectively. BamH I and Asc I are restriction enzyme recognition sites. The



nucleotide sequence of the vector is set forth in SEQ ID NO 4. The vector is digested with BamHI and Ascl to liberate the vector arms, which are used for ligation to the concatemer.

- 5       The ratio of vector arms to cassettes determines the maximum number of cassettes in the concatemer as illustrated in figure 8. The vector arms preferably are artificial chromosome vector arms such as those described in Fig. 7.

10       It is of course also possible to add stopper fragments to the concatenation solution, the stopper fragments each having a RS2 or RS2' in one end and a non-complementary overhang or a blunt end in the other end. The ratio of stopper fragments to cassettes can likewise control the maximum size of the concatemer.

15       The complete sequence of steps to be taken when starting with the isolation of mRNA until inserting into an entry vector may include the following steps

- 20           i)       isolating mRNA from an expression state,  
          ii)       obtaining substantially full length cDNA corresponding to the mRNA sequences,  
          iii)       inserting the substantially full length cDNA into a cloning site in a cassette in a primary vector, said cassette being of the general  
                    formula in 5'→3' direction:  
                    [RS1-RS2-SP-PR-CS-TR-SP-RS2'-RS1']  
                    wherein CS denotes a cloning site.

25       In preparation of the concatemer, genes may be isolated from different entry libraries to provide the desired selection of genes. Accordingly, concatenation may further comprise selection of vectors having expressible nucleotide sequences from at least two different expression states, such as from two different species. The two different species may be from two different classes, such as from two different  
30       divisions, more preferably from two different sub-kingdoms, such as from two different kingdoms.

As an alternative to including vector arms in the concatenation reaction it is possible to ligate the concatemer into an artificial chromosome selected from the group

comprising yeast artificial chromosome, mega yeast artificial chromosome, bacterial artificial chromosome, mouse artificial chromosome, human artificial chromosome.

Preferably at least one inserted concatemer further comprises a selectable marker.

5 The marker(s) are conveniently not included in the concatemer as such but rather in an artificial chromosome vector, into which the concatemer is inserted. Selectable markers generally provide a means to select, for growth, only those cells which contain a vector. Such markers are of two types: drug resistance and auxotrophy. A drug resistance marker enables cells to grow in the presence of an otherwise toxic  
10 compound. Auxotrophic markers allow cells to grow in media lacking an essential component by enabling cells to synthesise the essential component (usually an amino acid).

15 Illustrative and non-limiting examples of common compounds for which selectable markers are available with a brief description of their mode of action follow:

#### **Prokaryotic**

- Ampicillin: interferes with a terminal reaction in bacterial cell wall synthesis. The resistance gene (bla) encodes beta-lactamase which cleaves the beta-lactam ring of the antibiotic thus detoxifying it.
- 20 • Tetracycline: prevents bacterial protein synthesis by binding to the 30S ribosomal subunit. The resistance gene (tet) specifies a protein that modifies the bacterial membrane and prevents accumulation of the antibiotic in the cell.
- Kanamycin: binds to the 70S ribosomes and causes misreading of  
25 messenger RNA. The resistant gene (nptH) modifies the antibiotic and prevents interaction with the ribosome.
- Streptomycin: binds to the 30S ribosomal subunit, causing misreading of messenger RNA. The resistance gene (Sm) modifies the antibiotic and prevents interaction with the ribosome.
- 30 • Zeocin: this new bleomycin-family antibiotic intercalates into the DNA and cleaves it. The Zeocin resistance gene encodes a 13,665 dalton protein. This protein confers resistance to Zeocin by binding to the antibiotic and preventing it from binding DNA. Zeocin is effective on most aerobic cells and can be used for selection in mammalian cell lines, yeast, and bacteria.

**Eukaryotic**

- Hygromycin: a aminocyclitol that inhibits protein synthesis by disrupting ribosome translocation and promoting mistranslation. The resistance gene (hph) detoxifies hygromycin -B- phosphorylation.
- 5 • Histidinol: cytotoxic to mammalian cells by inhibiting histidyl-tRNA synthesis in histidine free media. The resistance gene (hisD) product inactivates histidinol toxicity by converting it to the essential amino acid, histidine.
- Neomycin (G418): blocks protein synthesis by interfering with ribosomal functions. The resistance gene ADH encodes amino glycoside  
10 phosphotransferase which detoxifies G418.
- Uracil: Laboratory yeast strains carrying a mutated gene which encodes orotidine -5'- phosphate decarboxylase, an enzyme essential for uracil biosynthesis, are unable to grow in the absence of exogenous uracil. A copy of the wild-type gene (ura4+, *S. pombe* or URA3 *S. cerevisiae*) carried on  
15 the vector will complement this defect in transformed cells.
- Adenosine: Laboratory strains carrying a deficiency in adenosine synthesis may be complemented by a vector carrying the wild type gene, ADE 2.
- Amino acids: Vectors carrying the wild-type genes for LEU2, TRP 1, HIS 3 or LYS 2 may be used to complement strains of yeast deficient in these genes.
- 20 • Zeocin: this new bleomycin-family antibiotic intercalates into the DNA and cleaves it. The Zeocin resistance gene encodes a 13,665 dalton protein. This protein confers resistance to Zeocin by binding to the antibiotic and preventing it from binding DNA. Zeocin is effective on most aerobic cells and can be used for selection in mammalian cell lines, yeast, and bacteria.

25

**Transgenic cells**

In one aspect of the invention, the concatemers comprising the multitude of cassettes are introduced into a host cell, in which the concatemers can be  
30 maintained and the expressible nucleotide sequences can be expressed in a co-ordinated way. The cassettes comprised in the concatemers may be isolated from the host cell and re-assembled due to their uniform structure with –preferably – concatemer restriction sites between the cassettes.

The host cells selected for this purpose are preferably cultivable under standard laboratory conditions using standard culture conditions, such as standard media and protocols. Preferably the host cells comprise a substantially stable cell line, in which the concatemers can be maintained for generations of cell division. Standard techniques for transformation of the host cells and in particular methods for insertion of artificial chromosomes into the host cells are known.

It is also of advantage if the host cells are capable of undergoing meiosis to perform sexual recombination. It is also advantageous that meiosis is controllable through external manipulations of the cell culture. One especially advantageous host cell type is one where the cells can be manipulated through external manipulations into different mating types.

The genome of a number of species have already been sequenced more or less completely and the sequences can be found in databases. The list of species for which the whole genome has been sequenced increases constantly. Preferably the host cell is selected from the group of species, for which the whole genome or essentially the whole genome has been sequenced. The host cell should preferably be selected from a species that is well described in the literature with respect to genetics, metabolism, physiology such as model organism used for genomics research.

The host organism should preferably be conditionally deficient in the abilities to undergo homologous recombination. The host organism should preferably have a codon usage similar to that of the donor organisms. Furthermore, in the case of genomic DNA, if eukaryotic donor organisms are used, it is preferable that the host organism has the ability to process the donor messenger RNA properly, e.g., splice out introns.

The host cells can be bacterial, archaeobacteria, or eukaryotic and can constitute a homogeneous cell line or mixed culture. Suitable cells include the bacterial and eukaryotic cell lines commonly used in genetic engineering and protein expression.

Preferred prokaryotic host organisms may include but are not limited to *Escherichia coli*, *Bacillus subtilis*, *B. licheniformis*, *B. cereus*, *Streptomyces lividans*,

Streptomyces coelicolor, Pseudomonas aeruginosa, Myxococcus xanthus. Rhodococcus, Streptomyces, Actinomycetes, Corynebacteria, Bacillus, Pseudomonas, Salmonella, and Erwinia. The complete genome sequences of E. coli and Bacillus subtilis are described by Blattner et al., Science 277, 1454-1462 (1997); Kunst et al., Nature 390, 249-256 (1997)).

Preferred eukaryotic host organisms are mammals, fish, insects, plants, algae and fungi.

Examples of mammalian cells include those from, e.g., monkey, mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Preferred mammalian host cells include but are not limited to those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa etc (see Kriegler M. in "Gene Transfer and Expression: A Laboratory Manual", New York, Freeman & Co. 1990), and stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, kidney, liver, muscle, and skin cells.

Examples of insect cells include baculo lepidoptera.

Examples of plant cells include maize, rice, wheat, cotton, soybean, and sugarcane. Plant cells such as those derived from Nicotiana and Arabidopsis are preferred

Examples of fungi include penicillium, aspergillus, such as Aspergillus nidulans, podospora, neurospora, such as Neurospora crassa, saccharomyces, such as Saccharomyces cerevisiae (budding yeast), Schizosaccharomyces, such as Schizosaccharomyces pombe (fission yeast), Pichia spp, such as Pichia pastoris, and Hansenula polymorpha (methylotropic yeasts).

In a preferred embodiment the host cell is a yeast cell, and an illustrative and not limiting list of suitable yeast host cells comprise: baker's yeast, Kluyveromyces marxianus, K. lactis, Candida utilis, Phaffia rhodozyma, Saccharomyces boulardii, Pichia pastoris, Hansenula polymorpha, Yarrowia lipolytica, Candida paraffinica, Schwanniomyces castellii, Pichia stipitis, Candida shehatae, Rhodotorula glutinis, Lipomyces lipofer, Cryptococcus curvatus, Candida spp. (e.g. C. palmioleophila),

Yarrowia lipolytica, Candida guilliermondii, Candida, Rhodotorula spp., Saccharomycopsis spp., Aureobasidium pullulans, Candida brumptii, Candida hydrocarbofumarica, Torulopsis, Candida tropicalis, Saccharomyces cerevisiae, Rhodotorula rubra, Candida flaveri, Eremothecium ashbyii, Pichia spp., Pichia pastoris, Kluyveromyces, Hansenula, Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola.

The choice of host will depend on a number of factors, depending on the intended use of the engineered host, including pathogenicity, substrate range, environmental hardiness, presence of key intermediates, ease of genetic manipulation, and likelihood of promiscuous transfer of genetic information to other organisms. Particularly advantageous hosts are E. coli, lactobacilli, Streptomyces, Actinomycetes, Saccharomyces and filamentous fungi.

In any one host cell it is possible to make all sorts of combinations of expressible nucleotide sequences from all possible sources. Furthermore, it is possible to make combinations of promoters and/or spacers and/or introns and/or terminators in combination with one and the same expressible nucleotide sequence.

Thus in any one cell there may be expressible nucleotide sequences from two different expression states. Furthermore, these two different expression states may be from one species or advantageously from two different species. Any one host cell may also comprise expressible nucleotide sequences from at least three species, such as from at least four, five, six, seven, eight, nine or ten species, or from more than 15 species such as from more than 20 species, for example from more than 30, 40 or 50 species, such as from more than 100 different species, for example from more than 300 different species, such as from more than 500 different species, for example from more than 1000 different species, thereby obtaining combinations of large numbers of expressible nucleotide sequences from a large number of species.

In this way potentially unlimited numbers of combinations of expressible nucleotide sequences can be combined across different expression states. These different expression states may represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera. The different species may be from at least two different phylae, such as from at least two different

classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.

5 Any two of these species may be from two different classes, such as from two different divisions, more preferably from two different sub-kingdoms, such as from two different kingdoms. Thus expressible nucleotide sequences may be combined from a eukaryot and a prokaryot into one and the same cell.

10 According to another embodiment of the invention, the expressible nucleotide sequences may be from one and the same expression state. The products of these sequences may interact with the products of the genes in the host cell and form new enzyme combinations leading to novel biochemical pathways. Furthermore, by putting the expressible nucleotide sequences under the control of a number of promoters it becomes possible to switch on and off groups of genes in a co-ordinated manner. By doing this with expressible nucleotide sequences from only one expression  
15 states, novel combinations of genes are also expressed.

The number of concatemers in one single cell may be at least one concatemer per cell, preferably at least 2 concatemers per cell, more preferably 3 per cell, such as 4  
20 per cell, more preferably 5 per cell, such as at least 5 per cell, for example at least 6 per cell, such as 7, 8, 9 or 10 per cell, for example more than 10 per cell. As described above, each concatemer may preferably comprise up to 1000 cassettes, and it is envisaged that one concatemer may comprise up to 2000 cassettes. By inserting up to 10 concatemers into one single cell, this cell may thus be enriched  
25 with up to 20,000 heterologous expressible genes, which under suitable conditions may be turned on and off by regulation of the regulatable promoters.

Often it is more preferable to provide cells having anywhere between 10 and 1000 heterologous genes, such as 20-900 heterologous genes, for example 30 to 800  
30 heterologous genes, such as 40 to 700 heterologous genes, for example 50 to 600 heterologous genes, such as from 60 to 300 heterologous genes or from 100 to 400 heterologous genes which are inserted as 2 to 4 artificial chromosomes each containing one concatemer of genes. The genes may advantageously be located on 1 to 10 such as from 2 to 5 different concatemers in the cells. Each concatemer may  
35 advantageously comprise from 10 to 1000 genes, such as from 10 to 750 genes,

such as from 10 to 500 genes, such as from 10 to 200 genes, such as from 20 to 100 genes, for example from 30 to 60 genes, or from 50 to 100 genes.

5 The concatemers may be inserted into the host cells according to any known transformation technique, preferably according to such transformation techniques that ensure stable and not transient transformation of the host cell. The concatemers may thus be inserted as an artificial chromosome which is replicated by the cells as they divide or they may be inserted into the chromosomes of the host cell. The concatemer may also be inserted in the form of a plasmid such as a plasmid vector, 10 a phage vector, a viral vector, a cosmid vector, that is replicated by the cells as they divide. Any combination of the three insertion methods is also possible. One or more concatemers may thus be integrated into the chromosome(s) of the host cell and one or more concatemers may be inserted as plasmids or artificial chromosomes. One or more concatemers may be inserted as artificial chromosomes and one or 15 more may be inserted into the same cell via a plasmid.

## Examples

### Example 1

20

In the examples 1-3 an Asc1 site was introduced into the EcoR1 site in pYAC4 (Sigma, Burke DT et al. 1987, Science vol 236, p 806), so that sticky ends match the Asc1 site( = RS2 in general formula of this patent) of the cassettes in pEVE vectors

### 25 **Preparation of EVACs (EVolvable Artificial Chromosomes) including size fractioning**

#### preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB (sigma) with a single colony of E. coli DH5 $\alpha$  containing pYAC4-Asc
- 30 2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
3. digest 100 $\mu$ g pYAC4-Asc w. BamH1 and Asc1
4. dephosphorylate fragments and heat inactivate phosphatase( 20 min, 80 C)
5. purify fragments(e.g. Qiaquick Gel Extraction Kit)
6. run 1 % agarose gel to estimate amount of fragment

35



**Preparation of expression cassettes**

1. take 100 µg of plasmid preparation from each of the following libraries
  - a) pMA-CAR
  - b) pCA-CAR
  - 5 c) Phaffia cDNA library
  - d) Carrot cDNA library
2. digest w. Srf1( 10 units/prep, 37C overnight)
3. dephosphorylate (10 units/prep, 37C, 2h)
4. heat inactivate 80C, 20 min
- 10 5. concentrate and change buffer (precipitation or ultra filtration),
6. digest w. Asc1. (10 units/prep, 37 C, overnight)
7. adjust volume of preps to 100 µL

preparation of EVACs

15

Different types of EVACs have been made by varying the ratio of the different libraries that goes into the ligation reaction.

	pMA-CAR	pCA-CAR	Phaffia cDNA	Carrot cDNA
<b>EVAC</b>				
<u>A</u>	<u>40%</u>	<u>40%</u>	<u>10%</u>	<u>10%</u>
<u>B</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>

- 20 1. add ~100 ng arms of pYAC4-Asc /100 µg of cassette mixture
2. concentrate to < 33.5 µL
3. add 2.5 units of T4 DNA-ligase + 4 µL 10x ligase buffer. Adjust to 40 µL
4. ligate 3 h, 16 C
5. stop reaction by adding 2 µL of 500 mM EDTA
- 25 6. bring reaction volume to 125 µL, add 25 µL loading mix, heat at 60C for 5 min
7. distribute evenly in 10 wells of a 1% LMP agarose gel
8. run pulsed field gel (CHEF III, 1% LMP agarose, ½ strength TBE (BioRad), angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s,
- 30 run time 30 h)

9. stain part of the gel that contains molecular weight markers + 1 sample lane for quality check
10. cut remaining 9 sample lanes corresponding to mw. 97 – 194 kb(fraction 1); 194 – 291 kb(fraction 2); 291-365 kb(fraction 3) from the gel
- 5 11. agarose gel in high NaCl agarose buffer . 1 U agarase / 100µg gel. 40C 3 h
12. concentrate preparation to < 20 µL
13. transform suitable yeast strain w. preparation using alkali/cation transformation
14. plate on selective minimal media plates
- 10 15. incubate 30 C for 4-5 days
16. pick colonies
17. analyse colonies

## Example 2

### 15 **Preparation of EVACs (EVolvable Artificial Chromosomes) with direct transformation**

#### preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB with a single colony of DH5α containing pYAC4-Asc
2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
- 20 3. digest 100µg pYAC4-Asc w. BamH1 and Asc1
4. dephosphorylate fragments and heat inactivate phosphatase( 20 min, 80 C)
5. purify fragments(e.g. Qiaquick Gel Extraction Kit)
1. run 1 % agarose gel to estimate amount of fragment

### 25 **Preparation of expression cassettes**

1. take 100 µg of plasmid preparation from each of the following libraries
  - e) pMA-CAR
  - f) pCA-CAR
  - g) Phaffia cDNA library
  - 30 h) Carrot cDNA library
2. digest w. Srf1( 10 units/prep, 37C overnight)
3. dephosphorylate (10 units/prep, 37C, 2h)
4. heat inactivate 80C, 20 min
5. concentrate and change buffer (precipitation or ultra filtration),

6. digest w. Asc1. (10 units/prep, 37 C, overnight)
7. adjust volume of preps to 100  $\mu$ L

#### preparation of EVACs

5

Different types of EVACs have been made by varying the ratio of the different libraries that goes into the ligation reaction.

10

	pMA-CAR	pCA-CAR	Phaffia cDNA	Carrot cDNA
<b>EVAC</b>				
<u>A</u>	<u>40%</u>	<u>40%</u>	<u>10%</u>	<u>10%</u>
<u>B</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>

1. concentrate to < 32  $\mu$ L
2. add 1 unit of T4 DNA-ligase + 4  $\mu$ L 10x ligase buffer. Adjust to 40  $\mu$ L
3. ligate 2 h, 16 C
- 15 4. stop reaction by adding 2  $\mu$ L of 500 mM EDTA, heat inactivate 60C, 20 min
5. bring reaction volume to 500  $\mu$ L with dH<sub>2</sub>O, concentrate to 30  $\mu$ L
6. add 10 U Asc1, 4  $\mu$ L 10X Asc1 buffer, bring to 40  $\mu$ L
7. incubate at 37C for 1h (alternatively 15 min 30 min)
8. heat inactivate 60C, 20 min
- 20 9. add 2  $\mu$ g YAC4-Asc arms, 1 U T4 DNA ligase, 10  $\mu$ L 10X ligase buffer, bring to 100  $\mu$ L
10. incubate ON, 16C
11. add water to 500  $\mu$ L
12. concentrate to 25  $\mu$ L
- 25 13. transform suitable yeast strain w. preparation using alkali/cation transformation or other suitable transformation method
14. plate on selective minimal media plates
5. incubate 30 C for 4-5 days
6. pick colonies
- 30 17. analyse colonies

**Example 3****Preparation of EVACs (EVolvable Artificial Chromosomes) (Small scale preparation)**

5

**Preparation of expression cassettes**

1. inoculate 5 ml of LB-medium (Sigma) with library inoculum corresponding to a 10+ fold representation of library. Grow overnight
2. make plasmid miniprep from 1.5 ml of culture (E.g. Qiaprep spin miniprep kit)
- 10 3. digest plasmid w. Srf 1
4. dephosphorylate fragments and heat inactivate phosphatase( 20 min, 80 C)
5. digest w. Asc1
6. run 1/10 of reaction in 1% agarose to estimate amount of fragment

15 preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB with a single colony of E. coli DH5 $\alpha$  containing pYAC4-Asc
2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
3. digest 100 $\mu$ g pYAC4-Asc w. BamH1 and Asc1
- 20 4. dephosphorylate fragments and heat inactivate phosphatase( 20 min, 80 C)
5. purify fragments(E.g. Qiaquick Gel Extraction Kit)
6. run 1 % agarose gel to estimate amount of fragment

preparation of EVACs

- 25 1. mix expression cassette fragments with YAC-arms so that cassette/arm ration is ~1000/1
2. if needed concentrate mixture(use e.g. Microcon YM30) so fragment concentration > 75 ng/ $\mu$ L reaction
3. add 1 U T4 DNA ligase, incubate 16C, 1-3 h . Stop reaction by adding 1  $\mu$ L of
- 30 500 mM EDTA
4. run pulsed field gel (CHEF III, 1% LMP agarose, ½ strength TBE, angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s, run time 30 h) Load sample in 2 lanes.
5. stain part of the gel that contains molecular weight markers
- 35 6. cut sample lanes corresponding to mw. 100 – 200 kb

7. agarase gel in high NaCl agarase buffer . 1 U agarase / 100 mg gel
8. concentrate preparation to < 20  $\mu$ L
9. transform suitable yeast strain w. preparation using electroporation
10. plate on selective minimal media plates
- 5 11. incubate 30 C for 4-5 days
12. pick colonies

#### Example 4: cDNA libraries used in the production of EVACs

- 10 1. *Daucus carota*, carrot root library:
  - Full length
  - Oligo dT primed, directional cDNA library
  - cDNA library made using a pool of 3 Evolva EVE 4, 5 & 8 vectors (Fig. 4, 5, 6)
  - Number of independent clones:  $41.6 \times 10^6$
- 15 • Average size: 0.9 – 2.9 kb
- Number of different genes present: 5000 -10000
  
2. *Xanthophyllomyces dendrorhous*, (yeast), hole organism library
  - Full length
  - 20 • Oligo dT primed, directional cDNA library
  - cDNA library made using a pool of 3 Evolva EVE 4, 5 & 8 vectors (Fig. 4, 5, 6)
  - Number of independent clones:  $48.0 \times 10^6$
  - Average size: 1.0 – 3.8 kb
  - Number of different genes present: 5000 -10000
- 25 3. Target carotenoid gene cDNA library
  - Full length and normalised
  - Directional cDNA cloning
  - Library made by cloning each gene individually in 2 Evolva EVE 4, 5 & 8 vectors
  - 30 (Fig. 4, 5, 6)
  - Number of different genes: 48
  - Species and genes used:
    - *Gentiana* sp., ggps, psy, pds, zds, lcy-b, lcy-e, bhy, zep
    - *Rhodobacter capsulatus*, idi, crtC, crtF

- *Erwinia uredovora*, crtE, crtB, crtI, crtY, crtZ
- *Nostoc anabaena*, zds
- *Synechococcus* PCC7942, pds
- *Erwinia herbicola*, crtE, crtB, crtI, crtY, crtZ
- 5      • *Staphylococcus aureus*, crtM, crtN
- *Xanthophyllomyces dendrorhous*, crtI, crtYb
- *Capsicum annuum*, ccs, crtL
- *Nicotiana tabacum*, crtL, bchy
- *Prochlorococcus* sp., lcy-b, lcy-e
- 10      • *Saccharomyces cerevisiae*, idi
- *Corynebacterium* sp., crtI, crtYe, crtYf, crtEb
- *Lycopersicon esculentum*, psy-1
- *Neurospora crassa*, al1

## 15      **Example 5: Transformation of EVACs**

### **Example 5a: Transformation**

1. Inoculate a single colony into 100 ml YPD broth and grow with aeration at 30°C to mid log,  $2 \times 10^6$  to  $2 \times 10^7$  cells/ml.
2. Spin to pellet cells at 400 x g for 5 minutes; discard supernatant.
- 20      3. Resuspend cells in a total of 9 ml TE, pH 7.5. Spin to pellet cells and discard supernatant.
4. Gently resuspend cells in 5 ml 0.1 M Lithium/Cesium Acetate solution, pH 7.5.
5. Incubate at 30°C for 1 hour with gentle shaking.
6. Spin at 400 x g for 5 minutes to pellet cells and discard supernatant.
- 25      7. Gently resuspend in 1 ml TE, pH 7.5. Cells are now ready for transformation.
8. In a 1.5 ml tube combine:
  - 100 µl yeast cells
  - 5 µl Carrier DNA (10 mg/ml)
  - 5 µl Histamine Solution
  - 30      • 1/5 of an EVAC preparation in a 10 µl volume (max). (One EVAC preparation is made of 100 µg of concatenation reaction mixture)
9. Gently mix and incubate at room temperature for 30 minutes.

10. In a separate tube, combine 0.8 ml 50% (w/v) PEG 4000 and 0.1 ml TE and 0.1 ml of 1 M LiAc for each transformation reaction. Add 1 ml of this PEG/TE/LiAc mix to each transformation reaction. Mix cells into solution with gentle pipetting.
11. Incubate at 30°C for 1 hour.
12. Heat shock at 42°C for 15 minutes; cool to 30°C.
13. Pellet cells in a microcentrifuge at high speed for 5 seconds and remove supernatant.
14. Resuspend in 200 µl of rich media and plate in appropriate selective media
15. Incubate at 30°C for 48-72 hours until transformant colonies appear.

#### **Example 5b: Transformation of EVACs using electroporation**

100 ml of YPD is inoculated with one yeast colony and grown to  $OD_{600} = 1.3$  to 1.5. The culture is harvested by centrifuging at  $4000 \times g$  and 4°C. The cells are resuspended in 16 ml sterile H<sub>2</sub>O. Add 2 ml 10 × TE buffer, pH 7.5 and swirl to mix. Add 2 ml 10 × lithium acetate solution (1 M, pH 7.5) and swirl to mix. Shake gently 45 min at 30°C. Add 1.0 ml 0.5 M DTE while swirling. Shake gently 15 min at 30°C. The yeast suspension is diluted to 100 ml with sterile water. The cells are washed and concentrated by centrifuging at  $4000 \times g$ , resuspending the pellet in 50 ml ice-cold sterile water, centrifuging at  $4000 \times g$ , resuspending the pellet in 5 ml ice-cold sterile water, centrifuging at  $4000 \times g$  and resuspending the pellet in 0.1 ml ice-cold sterile 1 M sorbitol. The electroporation was done using a *Bio-Rad Gene Pulser*. In a sterile 1.5-ml microcentrifuge tube 40 µl concentrated yeast cells were mixed with 5 µl 1:10 diluted EVAC preparation. The yeast-DNA mix is transferred to an ice-cold 0.2-cm-gap disposable electroporation cuvette and pulsed at 1.5 kV, 25 µF, 200 Ω. 1 ml ice-cold 1 M sorbitol is added to the cuvette to recover the yeast. Aliquots are spread on selective plates containing 1 M sorbitol. Incubate at 30°C until colonies appear.

#### **Example 6: Rare restriction enzymes with recognition sequence and cleavage points**

In this example, rare restriction enzymes are listed together with their recognition sequence and cleavage points. (^) indicates cleavage points 5'-3' sequence and ( ) indicates cleavage points in the complementary sequence.

W = A or T; N = A, C, G, or T

6a)		Unique, palindromic overhang
5	AscI	GG^CGCG_CC
	AsiSI	GCG_AT^CGC
	CciNI	GC^GGCC_GC
	CspBI	GC^GGCC_GC
	FseI	GG_CCGG^CC
10	MchAI	GC^GGCC_GC
	NotI	GC^GGCC_GC
	PacI	TTA_AT^TAA
	SbfI	CC_TGCA^GG
	SdaI	CC_TGCA^GG
15	SgfI	GCG_AT^CGC
	SgrAI	CR^CCGG_YG
	Sse232I	CG^CCGG_CG
	Sse8387I	CC_TGCA^GG
20	6b) No overhang	
	BstRZ246I	ATTT^AAAT
	BstSWI	ATTT^AAAT
25	MspSWI	ATTT^AAAT
	MssI	GTTT^AAAC
	PmeI	GTTT^AAAC
	SmiI	ATTT^AAAT
	SrfI	GCCC^GGGC
30	SwaI	ATTT^AAAT
6c)		Non-palindromic and/or variable overhang
35	AarI	CACCTGCNNNN^NNNN_
	AbeI	CC^TCA_GC
	AloI	^NNNNN_NNNNNNNGAACNNNNNNNTCCNNNNNNN_NNNNN^
	BaeI	^NNNNN_NNNNNNNNNNACNNNNGTAYCNNNNNNN_NNNNN^
	BbvCI	CC^TCA_GC
40	CpoI	CG^GWC_CG
	CspI	CG^GWC_CG
	Pfl127I	RG^GWC_CY
	PpiI	^NNNNN_NNNNNNNGAACNNNNNNCTCNNNNNNNN_NNNNN^
	PpuMI	RG^GWC_CY
45	PpuXI	RG^GWC_CY
	Psp5II	RG^GWC_CY
	PspPPI	RG^GWC_CY
	RsrII	CG^GWC_CG
	Rsr2I	CG^GWC_CG
50	SanDI	GG^GWC_CC
	SapI	GCTCTTCN^NNN_
	SdiI	GGCCN_NNN^NGGCC
	SexAI	A^CCWGG_T
	SfiI	GGCCN_NNN^NGGCC
55	Sse1825I	GG^GWC_CC
	Sse8647I	AG^GWC_CT
	VpaK32I	GCTCTTCN^NNN_



## 6d) Meganucleases

	I-Sce I	TAGGGATAA_CAGG^GTAAT
	I-Ceu I	ACGGTC_CTAA^GGTAG
5	I-Cre I	AAACGTC_GTGA^GACAGTTT
	I-Sce II	GGTC_ACCC^TGAAGTA
	I-Sce III	GTTTTGG_TAAC^TATTTAT
	Endo. Sce I	GATGCTGC_AGGC^ATAGGCTTGTTTA
	PI-Sce I	GG_GTGC^GGAGAA
10	PI-Psp I	TGGCAAACAGCTA_TTAT^GGGTATTATGGGT
	I-Ppo I	CTCTC_TTAA^GGTAG
	HO	TTTCCGC_AACA^GT
	I-Tev I	NN_NN^NNTCAGTAGATGTTTTTCTTGGTCTACCGTTT

15

More meganucleases have been identified, but their precise sequence of recognition has not been determined, see e.g. [www.meganuclease.com](http://www.meganuclease.com)

20 **Example 7: Concatemer size limitation experiments (use of stoppers)**

Materials used:

pYAC4 (Sigma. Burke et al. 1987, science, vol 236, p 806) was digested w. EcoR1 and BamH1 and dephosphorylated

25 pSE420 (invitrogen) was linearised using EcoR1 and used as the model fragment for concatenation.

T4 DNA ligase (Amersham-pharmacia biotech) was used for ligation according to manufacturers instructions.

30 Method: Fragments and arms were mixed in the ratios (concentrations are arbitrary units) indicated on figures 9a and 9b. Ligation was allowed to proceed for 1 h at 16°C. Reaction was stopped by the addition of 1 µL 500 mM EDTA. Products were analysed by standard agarose GE (1 % agarose, ½ strength TBE) or by PFGE (CHEF III, 1% LMP agarose, ½ strength TBE, angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s, run time 30 h)

35

The results are shown in Figure 9, wherein it is shown that the size of concatemers is proportional to the ratio of cassettes per YAC arms.

40 **Example 8: Integration of expression cassettes into artificial chromosomes**

Integration of expression cassettes into YAC12 was done essentially as done by Sears D.D., Hieter P., Simchen G., Genetics, 1994, 138, 1055-1065.

An Ascl site was introduced into the Bgl II site of the integration vectors pGS534 and pGS525.

A  $\beta$ -galactosidase gene, as well as crtE, crtB, crtI and crtY from *Erwinia Uredovora* were cloned into pEVE4. These expression cassettes were ligated into Ascl of the modified integration vectors pGS534 and pGS525.

Linearised pGS534 and pGS525 containing the expression cassettes were transformed into haploid yeast strains containing the appropriate target YAC which carries the Ade<sup>r</sup> gene. Red Ade<sup>r</sup>- transformants were selected (the parent host strain is red due to the ade2-101 mutation).

Additional confirmation of correct integration of the  $\beta$ -galactosidase expression cassette was done using a  $\beta$ -galactosidase assay.

**Example 9: Re-transformation of cells that already contain Artificial chromosomes to obtain at least 2 artificial chromosomes per cell**

Yeast strains containing YAC12, Sears D.D., Hieter P., Simchen G., Genetics, 1994, 138, 1055-1065 were transformed with EVACs following the protocol described in example 4a. The transformed cells were plated on plates that select for cells that contained both YAC12 and EVACs.

**Example 10: Example of different expression patterns "phenotypes" obtained using the same yeast clones under different expression conditions:**

Colonies were picked with a sterile toothpick and streaked sequentially onto plates corresponding to the four repressed and/or induced conditions (-Ura/-Trp, -Ura/-

Trp/-Met, -Ura/-Trp/+200  $\mu$ M  $\text{Cu}_2\text{SO}_4$ , -Ura/-Trp/-Met/+200  $\mu$ M  $\text{Cu}_2\text{SO}_4$ ). 20 mg adenin was added to the media to suppress the ochre phenotype.

**Claims**

1. A library comprising a collection of individual cells, the cells being denoted

5 cell<sub>1</sub>, cell<sub>2</sub>, ..., cell<sub>i</sub>, wherein  $i \geq 2$ ,

each cell comprising at least one concatemer of individual oligonucleotide cassettes, each concatemer comprising a nucleotide sequence of the following formula:

10

$[rs_2\text{-}SP\text{-}PR\text{-}X\text{-}TR\text{-}SP\text{-}rs_1]_n$

wherein

15

rs<sub>1</sub> and rs<sub>2</sub> together denote a restriction site,

SP denotes a spacer of at least two bases,

X denotes an expressible nucleotide sequence,

20

PR denotes a promoter, capable of regulating the expression of X in the cell,

TR denotes a terminator, and

$n \geq 2$ , and

wherein at least one concatemer of cell<sub>1</sub> is different from a concatemer of cell<sub>2</sub>.

25

2. The library according to claim 1, wherein a concatemer of each cell comprises at least a first cassette and a second cassette, said first cassette being different from said second cassette.

30

3. The library according to claim 1, wherein substantially all cassettes of a concatemer are different.

4. The library according to claim 1, wherein substantially all cells of the library are different.

35

5. The library according to claim 1, said library comprising a collection of sublibraries.
6. The library according to claim 5, wherein a sublibrary is a collection of individual cells having at least one phenotype in common.
7. The library according to claim 6, wherein the at least one phenotype is selected from the group comprising the ability to grow on unusual substrates, the ability to grow on sublethal concentration of toxins, the ability to grow at a high temperature, the ability to grow at a low temperature, the ability to grow at elevated osmolality, the ability to grow at low osmolality, the ability to grow at high salinity, the ability to grow at low salinity, the ability to grow at elevated metal concentrations, the ability to grow at high CO<sub>2</sub> concentrations, the ability to grow at low CO<sub>2</sub> concentrations, the ability to grow at high O<sub>2</sub> concentrations, the ability to grow at low O<sub>2</sub> concentrations, the ability to provide special spectral properties, the ability to provide a special colour, the ability to have a deviating GST activity, the ability to have a deviating P450 activity.
8. The library according to claim 6, wherein a sublibrary is a collection of individual cells, said cells having – for at least one identical expressible DNA sequence – different promoters.
9. The library according to claim 6, wherein a sublibrary is a collection of individual cells, each cell having - in at least one cassette of the concatemer - identical expressible DNA sequences.
10. The library according to claim 1, comprising at least 20 individual cells, such as at least 50 individual cells.
11. The library according to claim 1, comprising at least 100 individual cells, such as at least 1,000 cells, for example at least 10,000 cells such as at least 100,000 cells, for example at least 1,000,000 cells, such as at least 1,000,000,000.
12. The library according to claim 1, comprising a collection of cells from one species.

13. The library according to claim 12, wherein said species is selected from prokaryotic species or mutants thereof.
- 5 14. The library according to claim 13, wherein said prokaryotic species is selected from *Escherichia coli*, *Bacillus subtilis*, *Streptomyces lividans*, *Streptomyces coelicolor*, *Pseudomonas aeruginosa*, *Myxococcus xanthus*.
- 10 15. The library according to claim 14, wherein said wherein said prokaryotic species is selected from *E. coli*.
16. The library according to claim 1, wherein said species is selected from eukaryotic species or mutants thereof.
- 15 17. The library according to claim 16, wherein said wherein said eukaryotic species is selected from mammals, insects, vertebrates, plants, fungi, yeasts; filamentous ascomycetes such as *Neurospora crassa* and *Aspergillus nidulans*; plant cells such as those derived from *Nicotiana* and *Arabidopsis*; mammalian host cells such as those derived from humans, monkeys and rodents, such as
- 20 chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa.
18. The library according to claim 16, wherein said wherein said eukaryotic species is selected from yeast or mutants thereof.
- 25 19. The library according to claim 18, said yeast being selected from the group comprising budding yeast, *Kluyveromyces marxianus*, *K. lactis*, *Candida utilis*, *Phaffia rhodozyma*, *Saccharomyces boulardii*, *Pichia pastoris*, *Hansenula polymorpha*, *Yarrowia lipolytica*, *Candida paraffinica*, *Schwanniomyces castellii*, *Pichia stipitis*, *Candida shehatae*, *Rhodotorula glutinis*, *Lipomyces lipofer*,
- 30 *Cryptococcus curvatus*, *Candida* spp. (e.g. *C. palmiophila*), *Yarrowia lipolytica*, *Candida guilliermondii*, *Candida*, *Rhodotorula* spp., *Saccharomycopsis* spp., *Aureobasidium pullulans*, *Candida brumptii*, *Candida hydrocarbofumarica*, *Torulopsis*, *Candida tropicalis*, *Saccharomyces cerevisiae*, *Rhodotorula rubra*, *Candida flaveri*, *Eremothecium ashbyii*, *Pichia* spp., *Kluyveromyces*, *Hansenula*,

Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola, or mutants thereof.

- 5 20. The library according to any of the preceding claims, wherein substantially all  $rs_1$ - $rs_2$  sequences are recognised by the same restriction enzyme, more preferably wherein substantially all  $rs_1$ - $rs_2$  sequences are substantially identical.
- 10 21. The library according to any of the preceding claims, wherein n in at least one concatemer in at least one cell cell, more preferably wherein n in substantially all concatemers in substantially all cells is at least 10, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000.
- 15 22. The library according to any of the preceding claims, wherein at least one cell comprises, more preferably substantially all cells comprise 2 concatemers per cell, more preferably 3 per cell, such as 4 per cell.
- 20 23. The library according to any of the preceding claims, wherein at least one cassette in one cell comprises an intron between the promoter and the expressible nucleotide sequence, more preferably substantially all cassettes in substantially all cells comprise an intron between the promoter and the expressible nucleotide sequence.
- 25 24. The library according to claim 1, wherein the difference is a difference in the spacer sequence and/or the promoter, and/or the expressible nucleotide sequence and/or the intron and/or terminator sequence.
- 30 25. The library according to claim 24, wherein the different expressible nucleotide sequences come from the same or from different expression states.
- 35 26. The library according to claim 25, wherein the different expression states represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera.

27. The library according to claim 26, wherein the different species are from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.
28. The library according to any of the preceding claims, wherein the nucleotide sequence of at least one concatemer, preferably the nucleotide sequence from substantially all concatemers have been designed to minimise the level of repeat sequences in any one concatemer.
29. The library according to claim 28, wherein recombination within the expressible nucleotide sequences has been minimised.
30. The library according to any of the preceding claims, wherein at least one concatemer is ligated into a plasmid or into an integration vector, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.
31. The library according to claim 30, wherein at least one concatemer is integrated into the host genome.
32. The library according to claim 30, wherein at least one concatemer is integrated into an artificial chromosome in the host cell.
33. The library according to any of the preceding claims, wherein the restriction site comprises a rare restriction site, having at least 7 bases in the recognition sequence, more preferably at least 8 bases, such as at least 9 bases, for example at least 10 bases.
34. The library according to any of the preceding claims, wherein the restriction enzyme recognising the  $rs_1$   $rs_2$  restriction site produces sticky ends upon cleavage of a double stranded nucleotide sequence, preferably wherein the sticky ends have a pre-determined nucleotide sequence.



35. The library according to any of the preceding claims, wherein the spacer sequences together comprise at least 50 bases, such as at least 60 bases, for example at least 75 bases, such as at least 100 bases, for example at least 150 bases, such as at least 200 bases, for example at least 250 bases, such as at least 300 bases, for example at least 400 bases, such as at least 500 bases, such as at least 750 bases, for example at least 1000 bases, such as at least 1100 bases, for example at least 1200 bases, such as at least 1300 bases, for example at least 1400 bases, such as at least 1500 bases, for example at least 1600 bases, such as at least 1700 bases, for example at least 1800 bases, such as at least 1900 bases, for example at least 2000 bases, such as at least 2100 bases, for example at least 2200 bases, such as at least 2300 bases, for example at least 2400 bases, such as at least 2500 bases, for example at least 2600 bases, such as at least 2700 bases, for example at least 2800 bases, such as at least 2900 bases, for example at least 3000 bases, such as at least 3200 bases, for example at least 3500 bases, such as at least 3800 bases, for example at least 4000 bases, such as at least 4500 bases, for example at least 5000 bases, such as at least 6000 bases.

36. The library according to claim 35, wherein at least one of the spacer sequences comprises between 50 and 2500 bases, preferably between 100 and 2500 bases, preferably between 200 and 2300 bases, more preferably between 300 and 2100 bases, such as between 400 and 1900 bases, more preferably between 500 and 1700 bases, such as between 600 and 1500 bases, more preferably between 700 and 1400 bases.

37. A library comprising a collection of individual cells, the cells being denoted

$cell_1, cell_2, \dots, cell_i$ , wherein  $i \geq 2$ ,

each cell comprising at least two expression cassettes comprising a nucleotide sequence of the following formula:

$[rs_2-SP-PR-X-TR-SP-rs_1]$

wherein

rs<sub>1</sub> and rs<sub>2</sub> together denote a restriction site,

SP denotes a spacer of at least two bases,

5 PR denotes a promoter, capable of functioning in the cell,

X denotes an expressible nucleotide sequence,

TR denotes a terminator, and

10 wherein at least one of the expression cassettes comprises an expressible nucleotide sequence heterologous to the cell, and at least one of the cassettes of cell<sub>1</sub> is different from the cassettes of cell<sub>2</sub>.

38. The library according to claim 37, wherein substantially all cells of the library are different.

15 39. The library according to claim 37, said library comprising a collection of sublibraries.

40. The library according to claim 39, wherein a sublibrary is a collection of  
20 individual cells having at least one phenotype in common.

41. The library according to claim 40, wherein the at least one phenotype is selected from the group comprising the ability to grow on unusual substrates, the ability to grow on sublethal concentrations of toxins, the ability to grow at a high  
25 temperature, the ability to grow at a low temperature, the ability to grow at elevated osmolality, the ability to grow at low osmolality, the ability to grow at high salinity, the ability to grow at low salinity, the ability to grow at elevated metal concentrations, the ability to grow at high CO<sub>2</sub> concentrations, the ability to grow at low CO<sub>2</sub> concentrations, the ability to grow at high O<sub>2</sub> concentrations, the  
30 ability to grow at low O<sub>2</sub> concentrations, the ability to provide special spectral properties, the ability to provide a special colour, the ability to have a deviating GST activity, the ability to have a deviating P450 activity.

42. The library according to claim 37, wherein a sublibrary is a collection of individual cells, said cells having – for identical expressible DNA sequences – different promoters.
- 5      43. The library according to claim 42, wherein a sublibrary is a collection of individual cells, each cell having at least one cassette with identical expressible DNA sequences.
44. The library according to claim 37, comprising at least 20 individual cells.
- 10      45. The library according to claim 37, comprising at least 50 individual cells.
46. The library according to claim 37, comprising at least 100 individual cells, such as at least 1,000 cells, for example at least 10,000 cells such as at least 100,000
- 15      cells, for example at least 1,000,000 cells, such as at least 1,000,000,000.
47. The library according to claim 37, comprising a collection of cells from one species.
- 20      48. The library according to claim 47, wherein said species is selected from prokaryotic species or mutants thereof.
49. The library according to claim 48, wherein said prokaryotic species is selected from *Escherichia coli*, *Bacillus subtilis*, *Streptomyces lividans*, *Streptomyces*
- 25      *coelicolor* *Pseudomonas aeruginosa*, *Myxococcus xanthus*.
50. The library according to claim 49, wherein said wherein said prokaryotic species is selected from *E. coli*.
- 30      51. The library according to claim 47, wherein said species is selected from eukaryotic species or mutants thereof.
52. The library according to claim 51, wherein said wherein said eukaryotic species is selected from mammals, insects, vertebrates, plants, fungi, yeasts;
- 35      filamentous ascomycetes such as *Neurospora crassa* and *Aspergillus nidulans*;

plant cells such as those derived from *Nicotiana* and *Arabidopsis*; mammalian host cells such as those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa.

5 53. The library according to claim 51, wherein said wherein said eukaryotic species is selected from fungi or mutants thereof.

54. The library according to claim 53, said yeast being selected from the group comprising budding yeast, *Kluyveromyces marxianus*, *K. lactis*, *Candida utilis*,  
10 *Phaffia rhodozyma*, *Saccharomyces boulardii*, *Pichia pastoris*, *Hansenula polymorpha*, *Yarrowia lipolytica*, *Candida paraffinica*, *Schwanniomyces castellii*, *Pichia stipitis*, *Candida shehatae*, *Rhodotorula glutinis*, *Lipomyces lipofer*, *Cryptococcus curvatus*, *Candida* spp. (e.g. *C. palmiophila*), *Yarrowia lipolytica*, *Candida guilliermondii*, *Candida*, *Rhodotorula* spp., *Saccharomycopsis*  
15 spp., *Aureobasidium pullulans*, *Candida brumptii*, *Candida hydrocarbofumarica*, *Torulopsis*, *Candida tropicalis*, *Saccharomyces cerevisiae*, *Rhodotorula rubra*, *Candida flaveri*, *Eremothecium ashbyii*, *Pichia* spp., *Kluyveromyces*, *Hansenula*, *Kloeckera*, *Pichia*, *Pachysolen* spp., or *Torulopsis bombicola*, or mutants thereof.

20

55. The library according to any of the preceding claims 37 to 54, wherein substantially all  $rs_1$ - $rs_2$  sequences are recognised by the same restriction enzyme, more preferably wherein substantially all  $rs_1$ - $rs_2$  sequences are substantially identical.

25

56. The library according to any of the preceding claims 37 to 55, wherein at least one cell comprises, more preferably wherein substantially all cells comprise at least 10 cassettes, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at  
30 least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000, such as at least 2500, for example at least 5000, such as at least 7500, for example at least 10,000.

57. The library according to any of the preceding claims 37 to 56, wherein at least one cassette in one cell comprises an intron capable of being identified and treated as an intron in the host cell between the promoter and the expressible nucleotide sequence, more preferably substantially all cassettes in substantially all cells comprise an intron between the promoter and the expressible nucleotide sequence.
58. The library according to claim 37, wherein the difference is a difference in the spacer sequence and/or the promoter, and/or the expressible nucleotide sequence and/or the intron and/or terminator sequence.
59. The library according to claim 58, wherein the different expressible nucleotide sequences come from the same or from different expression states.
60. The library according to claim 59, wherein the different expression states represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera.
61. The library according to claim 60, wherein the different species are from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different subkingdoms, such as from at least two different kingdoms.
62. The library according to any of the preceding claims 37 to 61, wherein the nucleotide sequence of at least one cassette, preferably the nucleotide sequence from substantially all cassettes have been designed to minimise the level of repeat sequences in any one cassettes.
63. The library according to any of the preceding claims 37 to 62, wherein recombination within the expressible nucleotide sequences has been minimised.
64. The library according to any of the preceding claims 37 to 63, wherein at least one cassette is ligated into a plasmid or into an integration vector, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.

65. The library according to any of the preceding claims 37 to 64, wherein at least one cassette is integrated into the host genome.

5 66. The library according to any of the preceding claims 37 to 65, wherein at least one cassette is integrated into an artificial chromosome in the host cell.

67. A library comprising a collection of individual cells, the cells being denoted

10  $cell_1, cell_2, \dots, cell_i$ , wherein  $i \geq 2$ ,

each cell comprising a random combination of heterologous oligonucleotides having the general formula:

15 [PR-X]

wherein

X denotes an expressible nucleotide sequence, and

20 PR denotes an independently controllable promoter being operably associated with X.

68. The library according to claim 67, wherein the random combinations are made from a two dimensional array of promoters and heterologous expressible nucleotide sequences.

25 69. The library according to any of the preceding claims 67 to 68, wherein each cell comprises an individual selection of combinations of promoters and heterologous expressible nucleotide sequences drawn individually from the same pool of promoters and heterologous expressible nucleotide sequences.

30 70. The library according to any of the preceding claims 67 to 69, wherein the library comprises at least 2 different independently controllable promoters, such as at least 3, for example at least 4, such as at least 5, for example at least 6, such as at least 7, for example at least 8, such as at least 9, for example at least 10,

such as at least 15, for example at least 25, such as at least 50, for example at least 75, such as at least 100.

5 71. The library according to any of the preceding claims 67 to 70, comprising an externally controllable promoter.

72. The library according to claim 71, comprising an inducible and/or a repressible promoter.

10 73. The library according to claim 71, comprising at least one promoter comprising both repressible and inducible elements.

15 74. The library according to any of the preceding claims 67 to 73, comprising at least one promoter being chemically inducible and/or repressible and/or inducible/repressible by temperature, and/or inducible/repressible according to mating type and/or inducible/repressible according to physical factors and/or inducible repressible according to environmental factors.

20 75. The library according to any of the preceding claims 67 to 74, wherein at least one promoter is induced by any factor selected from the group comprising carbohydrates; galactose; low inorganic phosphate levels; temperature; gaseous environment; pressure; pH; low or high temperature shift; metals or metal ions; copper ions; hormones; dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage; developmental stage; induced in MAT $\alpha$  cells; synthetic inducers; gal inducer.

25

30 76. The library according to any of the preceding claims 67 to 75, wherein at least one promoter is repressed by any factor selected from the group comprising carbohydrates; galactose; low inorganic phosphate levels; temperature; low or high temperature shift; metals or metal ions; copper ions; hormones; dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; gaseous environment; pressure; pH; growth stage; developmental stage; induced in MAT $\alpha$  cells; synthetic inducers; gal inducer; high inorganic phosphate levels; methionine; glycerol; repressed in MAT $\alpha$  or  $\alpha/\alpha$  cells

35

77. The library according to any of the preceding claims 67 to 76, comprising at least one promoter selected from the group comprising ADH 1, PGK 1, GAP 491, TPI, PYK, ENO, PMA 1, PHO5, GAL 1, GAL 2, GAL 10, MET25, ADH2, MEL 1, CUP 1, HSE, MF $\alpha$  1/Mfa 1, AOX, MOX, SV40, CaMV, Opaque-2, GRE, ARE, PGK/ARE hybrid, CYC/GRE hybrid, TPI/ $\alpha$ 2 operator, AOX 1, MOX A.
78. The library according to claim 67, wherein at least one promoter is a synthetic promoter.
79. The library according to any of the preceding claims 67 to 78, wherein at least one heterologous expressible nucleotide sequence is found in at least 2 cells, such as at least 3 cells, for example at least 5 cells, such as at least 10 cells, for example at least 25 cell, such as at least 50 cells, for example at least 100 cells, such as at least 500 cells, for example at least 1000 cells.
80. The library according to any of the preceding claims 67 to 79, wherein at least one cell comprises a group of heterologous expressible nucleotide sequences under the control of a first promoter, the group comprising at least 5 heterologous expressible nucleotide sequences, such as at least 10 heterologous expressible nucleotide sequences, for example at least 15 heterologous expressible nucleotide sequences, such as at least 25 heterologous expressible nucleotide sequences, for example at least 50 heterologous expressible nucleotide sequences, such as at least 75 heterologous expressible nucleotide sequences, for example at least 100 heterologous expressible nucleotide sequences, such as at least 250 heterologous expressible nucleotide sequences, for example at least 500 heterologous expressible nucleotide sequences.
81. The library according to claim 80, wherein a cell comprises at least a second group of heterologous expressible nucleotide sequences under the independent control of second promoter, such as at least a third group of heterologous expressible nucleotide sequences under the independent control of a third promoter, for example at least a fourth group of heterologous expressible nucleotide sequences under the independent control of a fourth promoter, such as at least a fifth group of heterologous expressible nucleotide sequences under



the independent control of a fifth promoter, for example at least a sixth group of heterologous expressible nucleotide sequences under the independent control of a sixth promoter, such as at least a seventh group of heterologous expressible nucleotide sequences under the independent control of a seventh promoter, such as at least a eighth group of heterologous expressible nucleotide sequences under the independent control of a eighth promoter, for example at least a ninth group of heterologous expressible nucleotide sequences under the independent control of a ninth promoter, such as at least a tenth group of heterologous expressible nucleotide sequences under the independent control of a tenth promoter.

82. The library according to any of the preceding claims 67 to 81, wherein the different expressible nucleotide sequences come from the same or from different expression states.

83. The library according to claim 82, wherein the different expression states represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera.

84. The library according to claim 83, wherein the different species are from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different subkingdoms, such as from at least two different kingdoms.

85. The library according to any of the preceding claims 67 to 84, wherein the nucleotide sequence of at least one cassette, preferably the nucleotide sequence from substantially all cassettes have been designed to minimise the level of repeat sequences in any one cassettes.

86. The library according to claim 85, wherein recombination within the expressible nucleotide sequences has been minimised.

87. The library according to any of the preceding claims 67 to 86, wherein at least one cassette is ligated into a plasmid or into an integration vector, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.

88. The library according to any of the preceding claims 67 to 87, wherein at least one cassette is integrated into the host genome.

5 89. The library according to any of the preceding claims 67 to 88, wherein at least one cassette is integrated into an artificial chromosome in the host cell.

90. The library according to any of the preceding claims 67 to 89, said library comprising a collection of sub-libraries.

10

91. The library according to claim 90, wherein a sub-library is a collection of individual cells having at least one phenotype in common.

15

92. The library according to claim 91, wherein the at least one phenotype is selected from the group comprising the ability to grow on unusual substrates, the ability to grow on sublethal concentrations of toxins, the ability to grow at a high temperature, the ability to grow at a low temperature, the ability to grow at elevated osmolality, the ability to grow at low osmolality, the ability to grow at high salinity, the ability to grow at low salinity, the ability to grow at elevated metal concentrations, the ability to grow at high CO<sub>2</sub> concentrations, the ability to grow at low CO<sub>2</sub> concentrations, the ability to grow at high O<sub>2</sub> concentrations, the ability to grow at low O<sub>2</sub> concentrations, the ability to provide special spectral properties, the ability to provide a special colour, the ability to have a deviating GST activity, the ability to have a deviating P450 activity.

25

93. The library according to claim 90, wherein a sublibrary is a collection of individual cells, said cells having – for at least one identical expressible DNA sequence, more preferably for substantially all identical expressible nucleotide sequences – different promoters.

30

94. The library according to claim 90, wherein a sublibrary is a collection of individual cells, each cell having - in at least one cassette of the concatemer - identical expressible DNA sequences.

95. The library according to claim 67, comprising at least 20 individual cells, such as at least 50 individual cells.
96. The library according to claim 67, comprising at least 100 individual cells, such as at least 1,000 cells, for example at least 10,000 cells such as at least 100,000 cells, for example at least 1,000,000 cells, such as at least 1,000,000,000 cells.
97. The library according to claim 67, comprising a collection of cells from one species.
98. The library according to claim 97, wherein said species is selected from prokaryotic species or mutants thereof.
99. The library according to claim 98, wherein said prokaryotic species is selected from *Escherichia coli*, *Bacillus subtilis*, *Streptomyces lividans*, *Streptomyces coelicolor*, *Pseudomonas aeruginosa*, *Myxococcus xanthus*.
100. The library according to claim 97, wherein said species is selected from eukaryotic species or mutants thereof.
101. The library according to claim 100, wherein said wherein said eukaryotic species is selected from mammals, insects, vertebrates, plants, fungi, yeasts; filamentous ascomycetes such as *Neurospora crassa* and *Aspergillus nidulans*; plant cells such as those derived from *Nicotiana* and *Arabidopsis*; mammalian host cells such as those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa.
102. The library according to claim 100, wherein said wherein said eukaryotic species is selected from fungi or mutants thereof.
103. The library according to claim 102, said yeast being selected from the group comprising budding yeast, *Kluyveromyces marxianus*, *K. lactis*, *Candida utilis*, *Phaffia rhodozyma*, *Saccharomyces boulardii*, *Pichia pastoris*, *Hansenula polymorpha*, *Yarrowia lipolytica*, *Candida paraffinica*, *Schwanniomyces castellii*,

Pichia stipitis, Candida shehatae, Rhodotorula glutinis, Lipomyces lipofer, Cryptococcus curvatus, Candida spp. (e.g. C. palmioleophila), Yarrowia lipolytica, Candida guilliermondii, Candida, Rhodotorula spp., Saccharomycopsis spp., Aureobasidium pullulans, Candida brumptii, Candida hydrocarbofumarica, Torulopsis, Candida tropicalis, Saccharomyces cerevisiae, Rhodotorula rubra, Candida flaveri, Eremothecium ashbyii, Pichia spp., Kluyveromyces, Hansenula, Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola, or mutants thereof.

104. A library comprising at least one library or at least one sub-library as defined in any of claims 1 to 103.

105. A method of producing a library comprising a collection of individual cells, comprising the steps:

- i) providing a population of nucleotide cassettes having the general formula  $[rs_2\text{-}SP\text{-}PR\text{-}X\text{-}TR\text{-}SP\text{-}rs_1]$ , wherein  $rs_1$  and  $rs_2$  together denote a restriction site, SP denotes a spacer of at least two bases, X denotes an expressible nucleotide sequence, PR denotes a promoter, capable of regulating the expression of X in the cell, TR denotes a terminator, and
- ii) assembling random sub-sets of the cassettes into concatemers comprising at least two cassettes,
- iii) ligating the concatemers into vectors,
- iv) introducing vectors into host cells,
- v) mixing at least two cells so that at least one concatemer of a first cell comprises a random sub-set of cassettes being different from a random sub-set of cassettes of a concatemer of a second cell.

106. The method according to claim 105, whereby the vectors comprise a plasmid or an integration vector, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.

107. The method according to claim 106, whereby the vectors comprise an artificial chromosome.

108. A method of producing a library comprising a collection of individual cells, comprising the steps:

- i) inserting at least two expressible nucleotides into the cloning site of at least two primary vectors comprising a cassette, the cassette comprising a nucleotide sequence of the general formula in 5'→3' direction: [RS1-RS2-SP-PR-CS-TR-SP-RS2-RS1'] wherein RS1 and RS1' denote restriction sites, RS2 denote a restriction site different from RS1 and RS1', SP denotes a spacer sequence of at least two nucleotides, PR denotes a promoter, CS denotes a cloning site, and TR denotes a terminator.
- ii) excising the cassettes using at least a restriction enzyme specific for RS1, RS1' RS2 and RS2' obtaining expression cassettes having the general formula [rs<sub>2</sub>-SP-PR-X-TR-SP-rs<sub>1</sub>], wherein rs<sub>1</sub> and rs<sub>2</sub> together denote a restriction site, and wherein X denotes an expressible nucleotide sequence,
- iii) inserting the expression cassettes into a vector,
- iv) transferring the expression cassettes into at least two host cells, and
- v) mixing at least two host cells having different cassettes.

109. A method of producing a library comprising a collection of individual cells, comprising the steps:

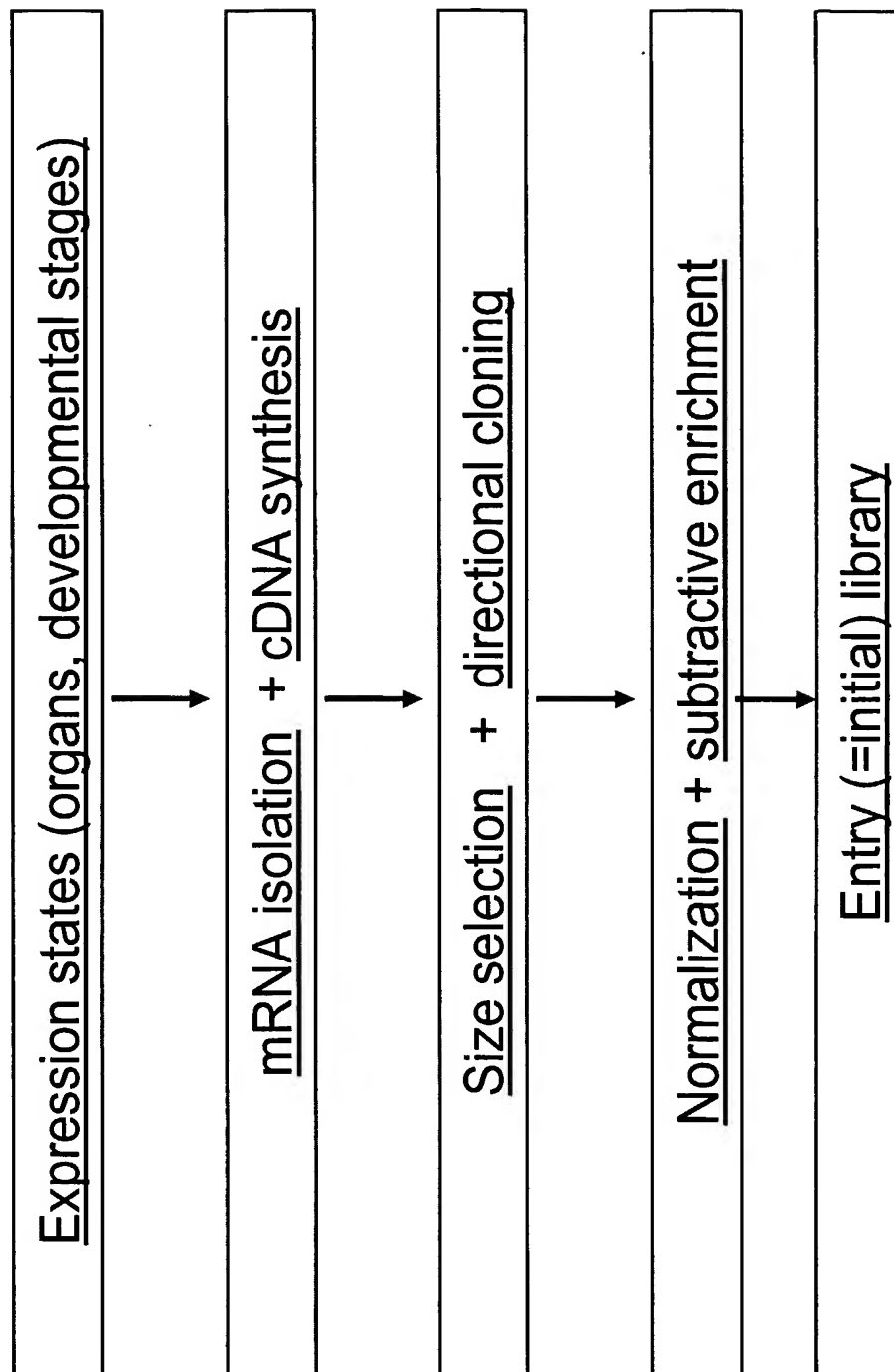
- i) providing at least one expressible nucleotide sequence,
- ii) ligating at least one expressible nucleotide sequence to a controllable promoter capable of functioning in a host cell obtaining a first expression construct,
- iii) ligating at least one expressible nucleotide sequence to another independently controllable promoter capable of functioning in a host cell, obtaining a second expression construct,
- iv) inserting constructs of step ii) and iii) into at least two host cells,
- v) mixing at least two cells having a different combination of independently controllable promoter and expressible nucleotide sequences.

110. The method according to claim 109, wherein expression constructs are concatenated prior to insertion into host cells.

111. An expression library obtainable by the method of any of the claims 105 to 110.

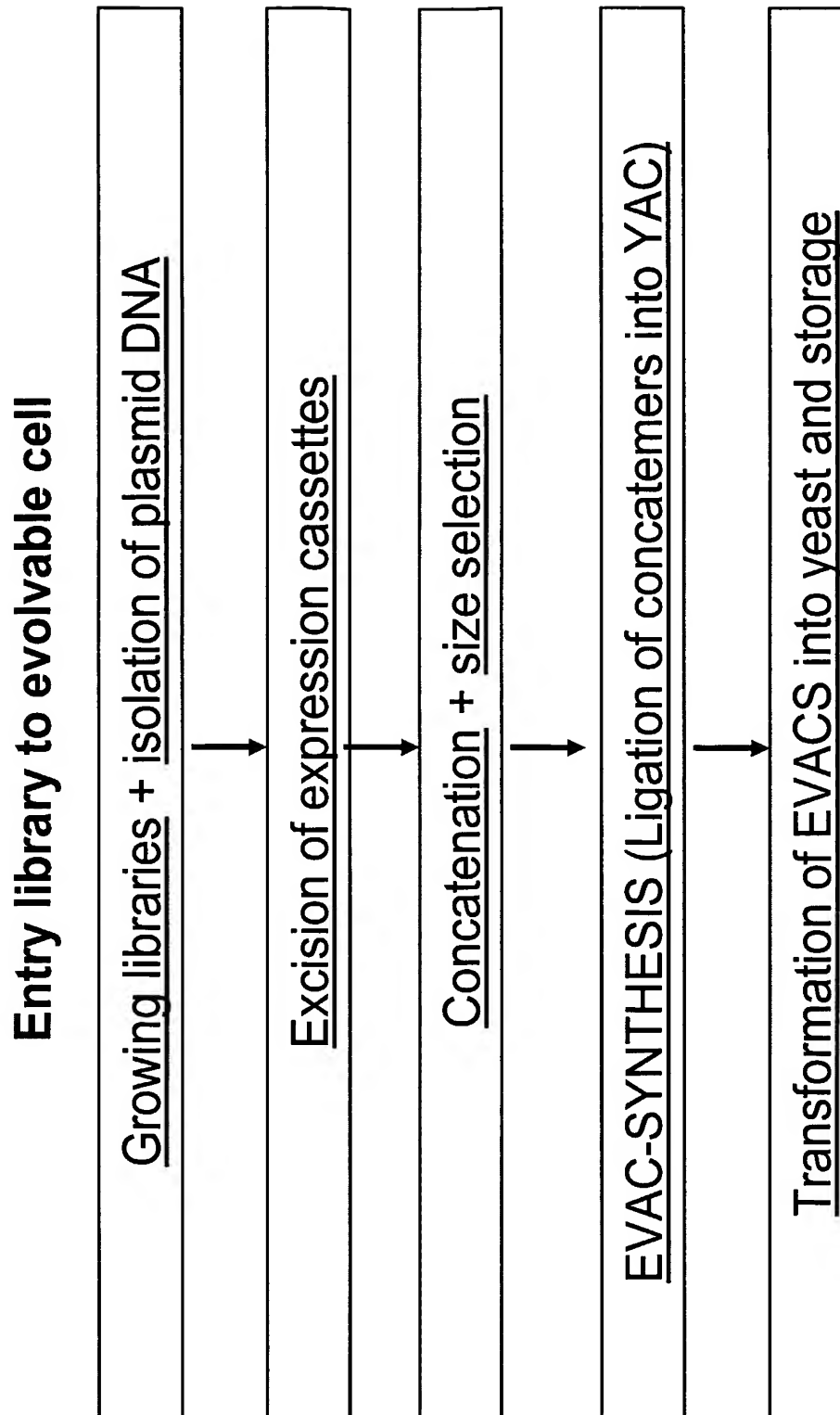
Fig. 1

## Constructing entry libraries



2/13

Fig. 2a





3/13

Fig. 2b

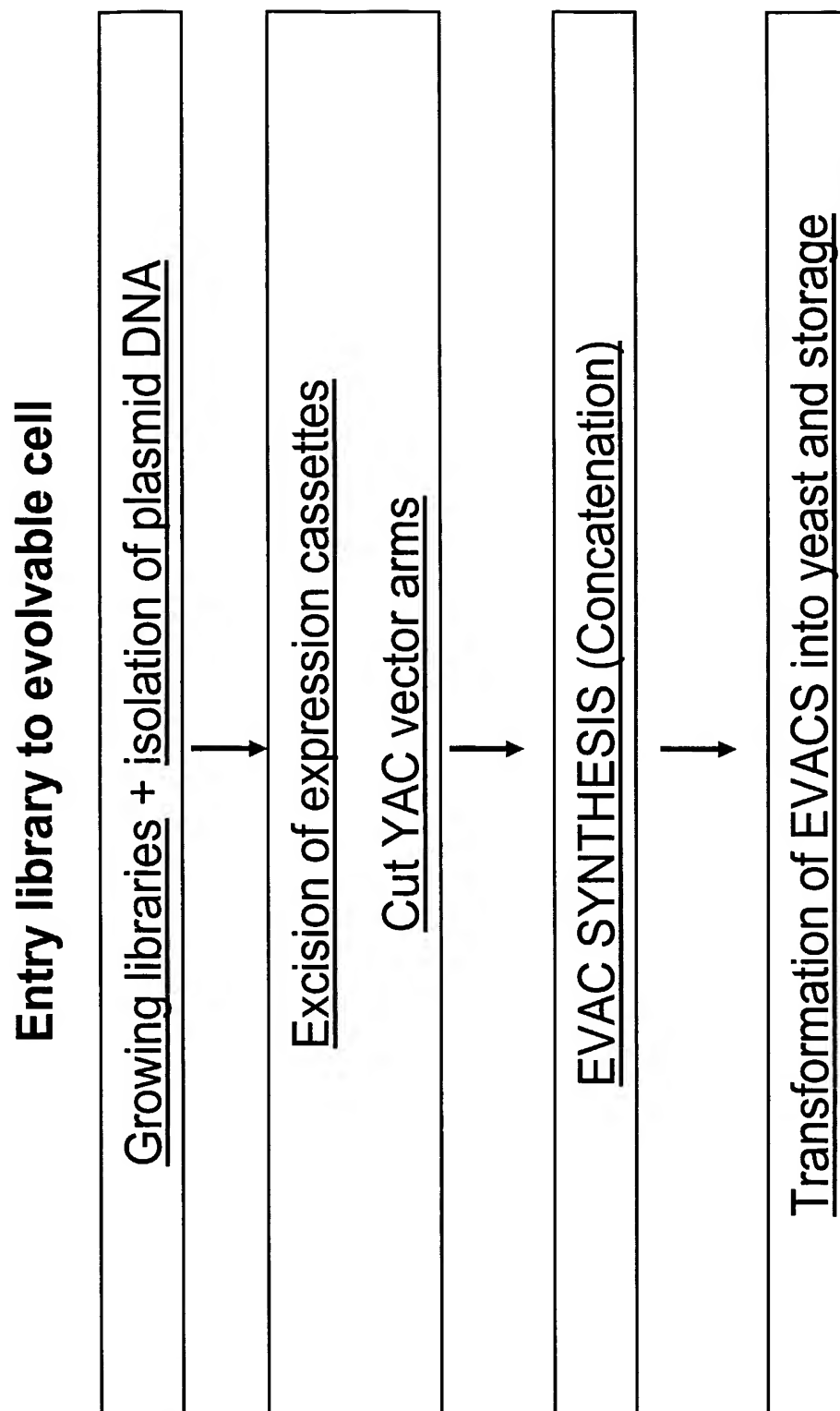


Fig. 3

## Model Entry Vector

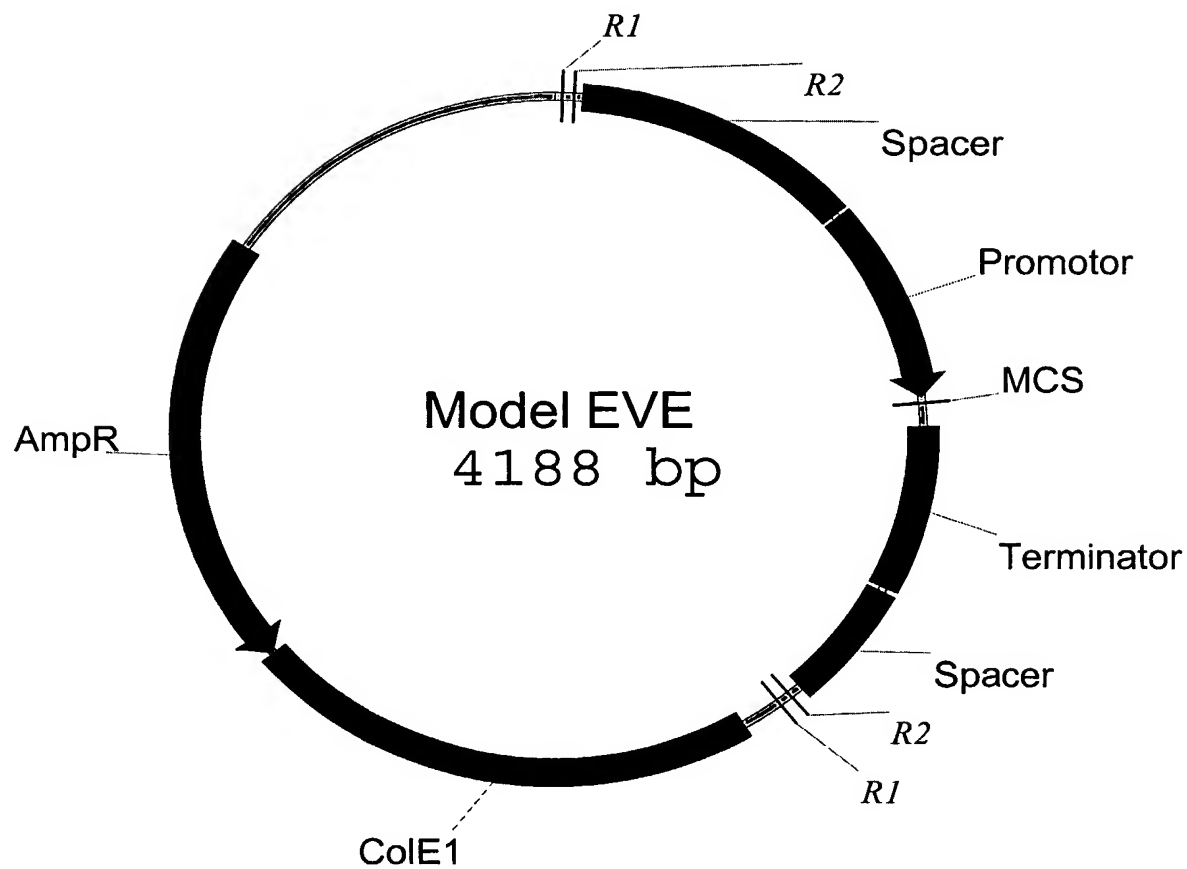
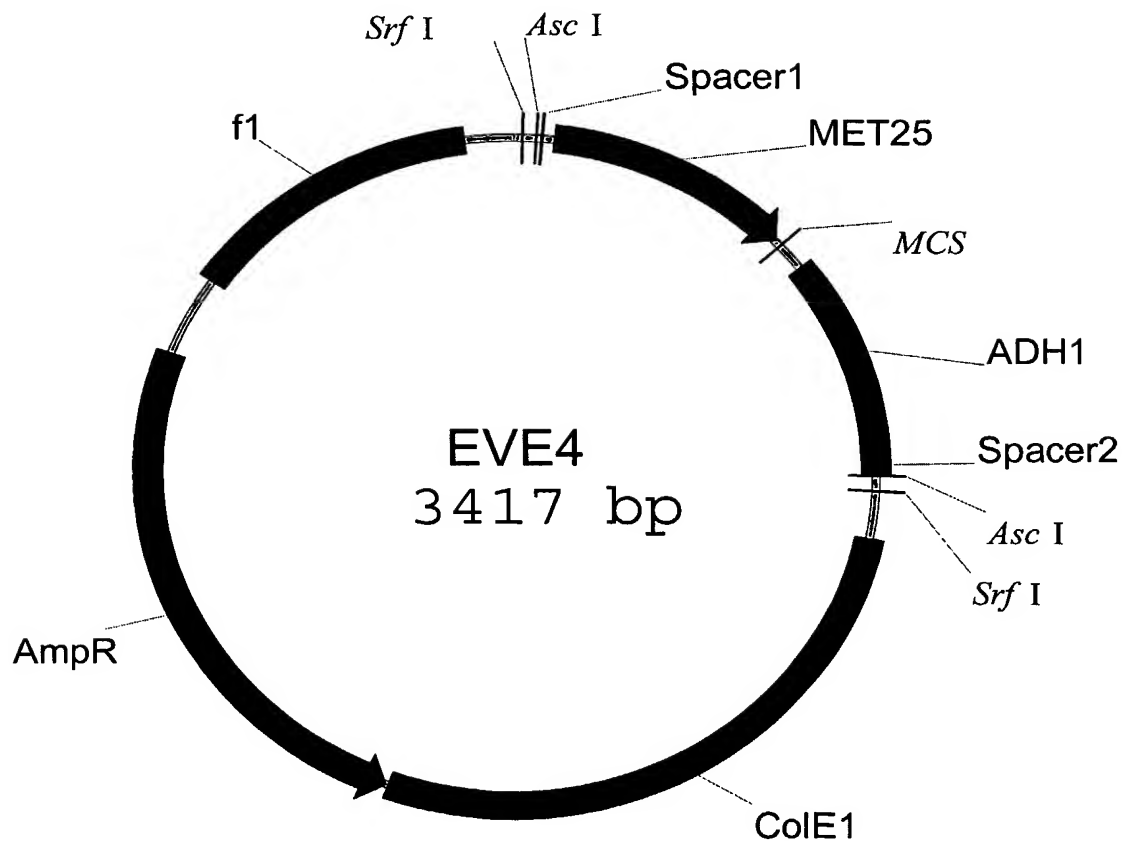


Fig. 4

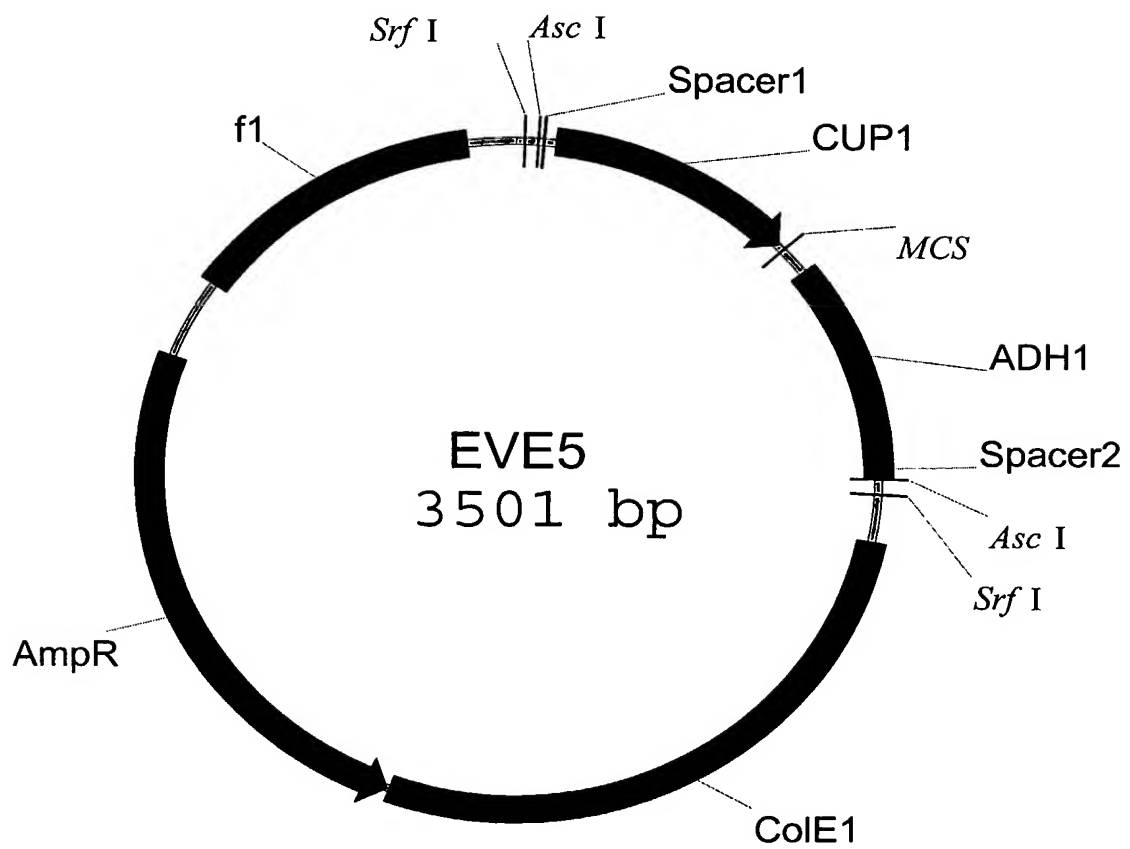
## EVE4 entry vector



6/13

Fig. 5

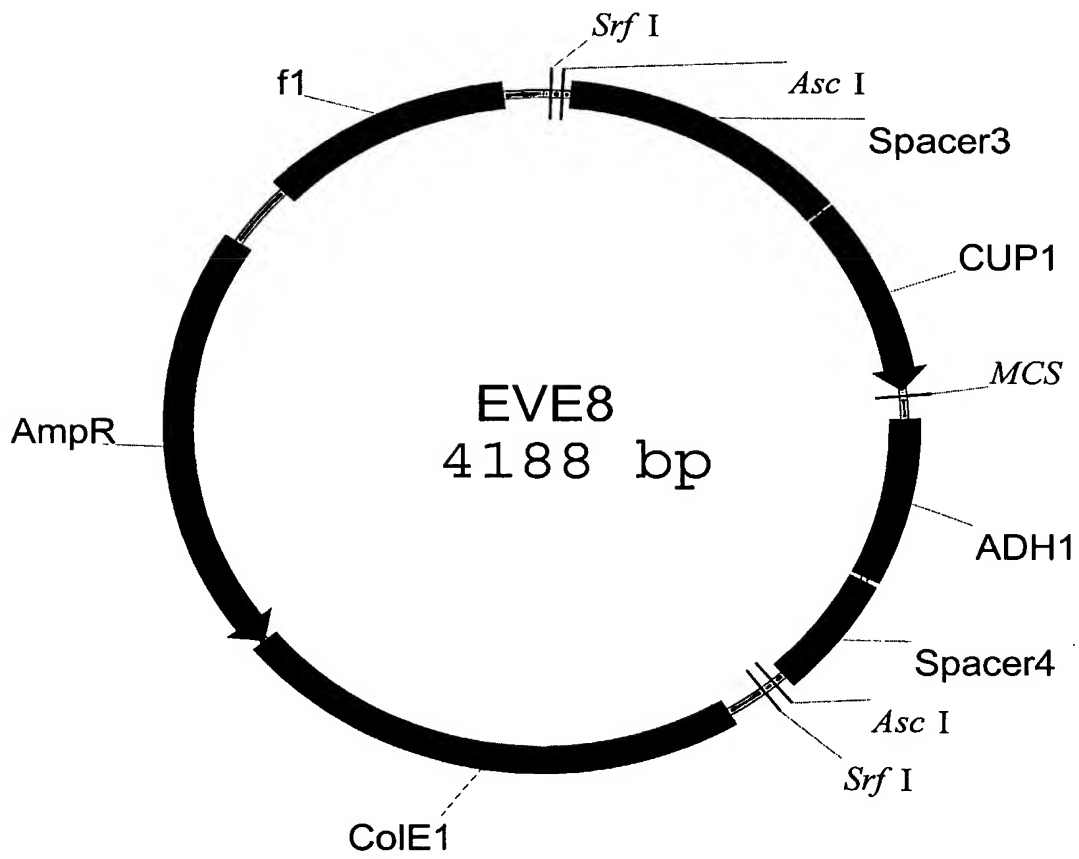
## EVE5 entry vector



7/13

Fig. 6

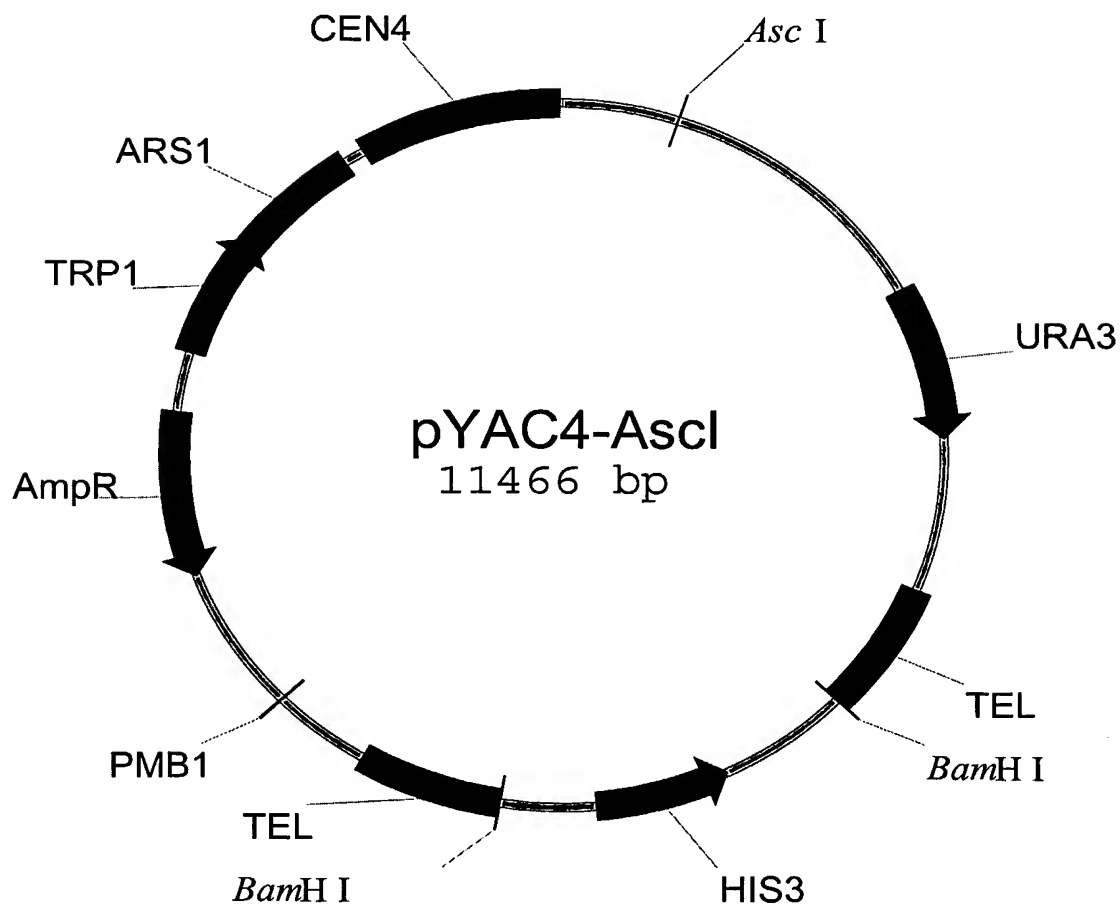
## EVE8 entry vector



8/13

Fig. 7

**pYAC4-AscI**  
Vector for providing EVACS arms



9/13

Fig. 8

# Synthesis of Concatemers

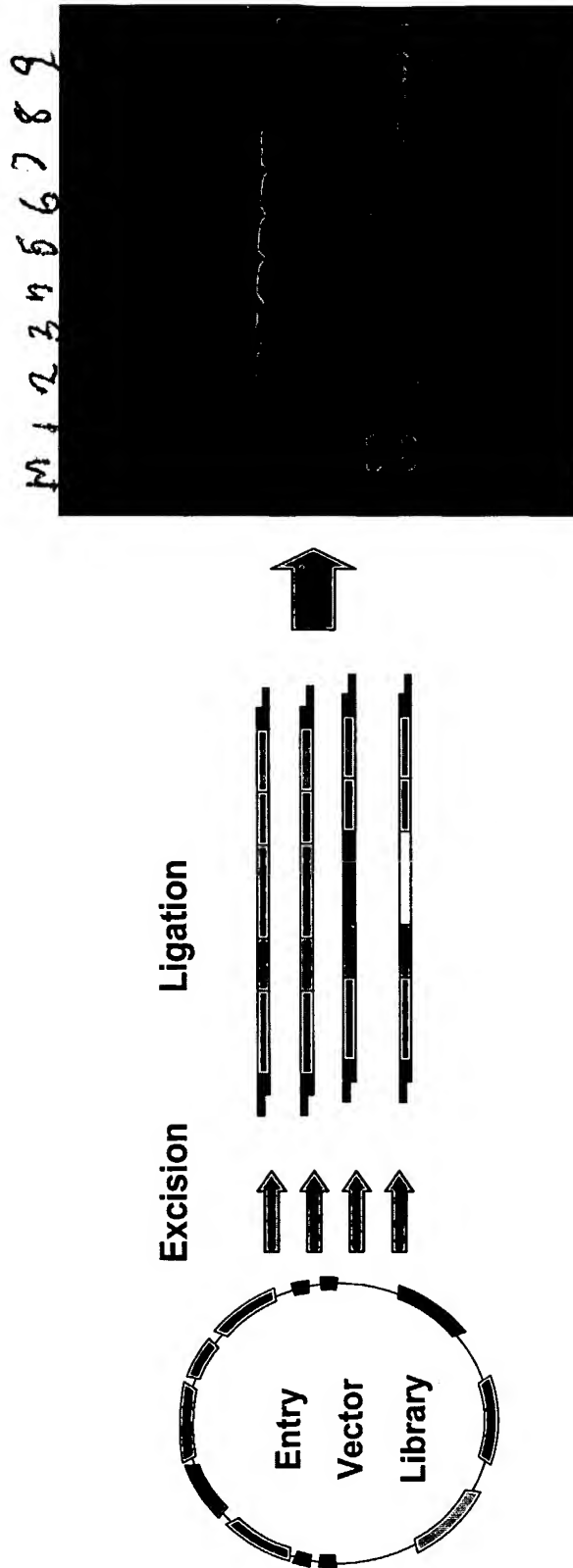
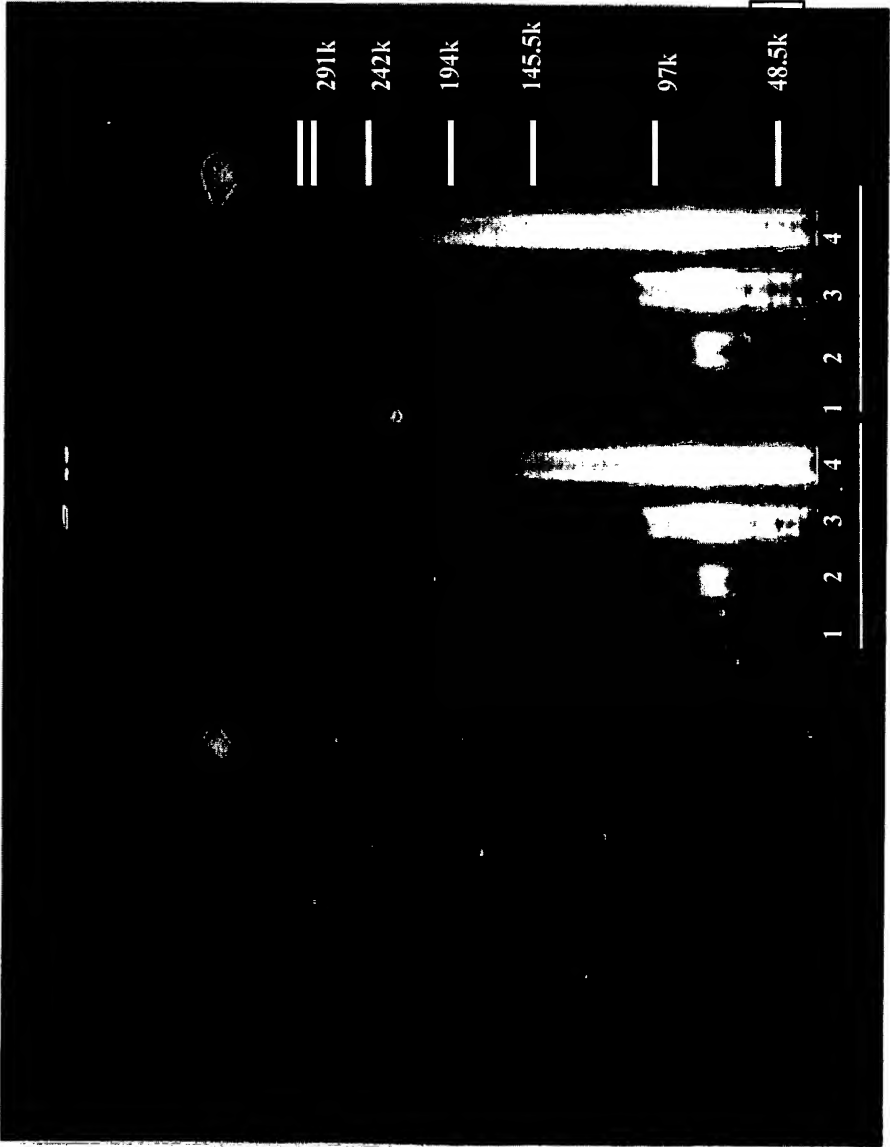


Fig. 9a



A: F/Y = 100, B: F/Y= 1000

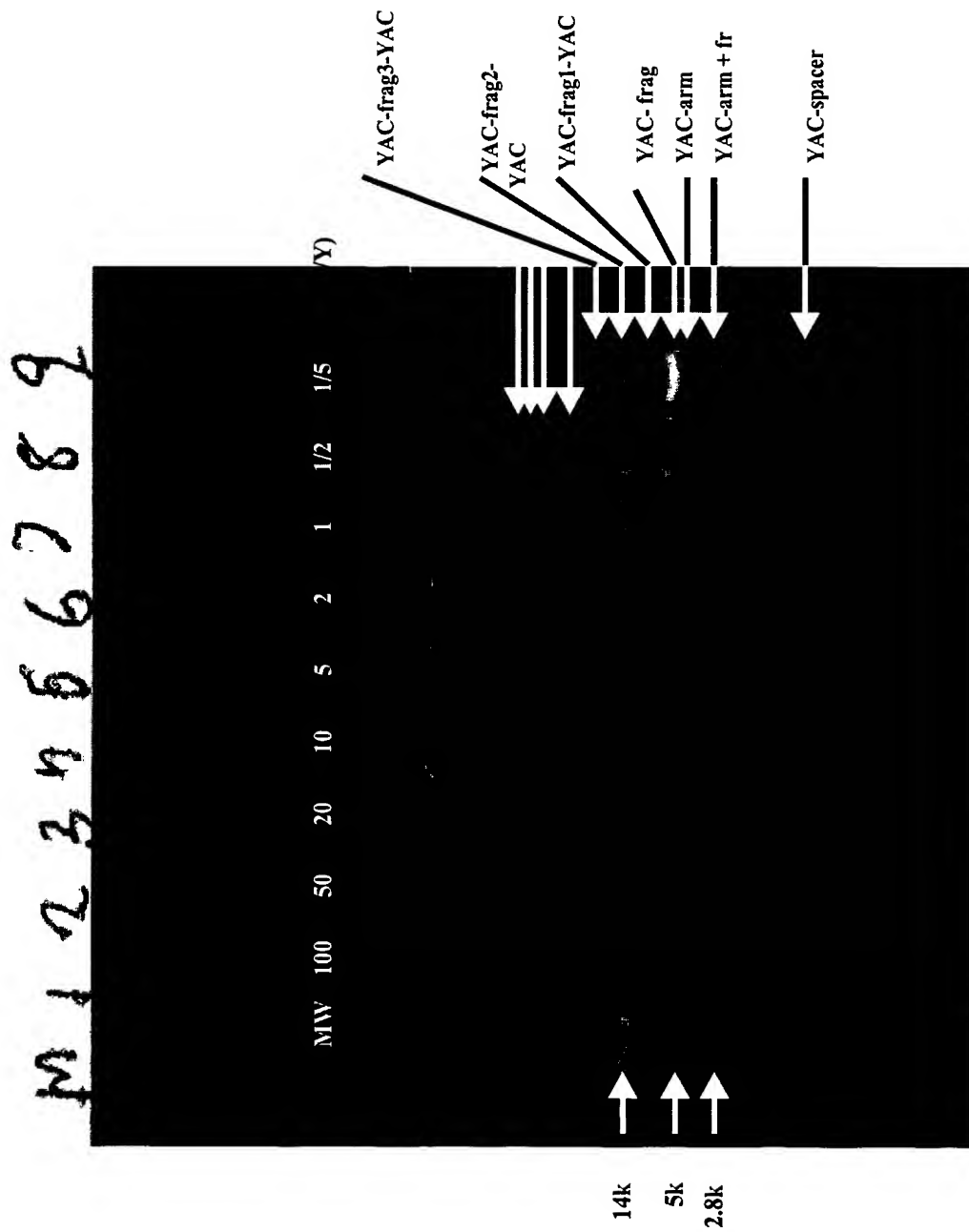
1: fragment conc.= 1  
2: fragment conc.= 2  
3: fragment conc.= 5  
4: fragment conc.= 10

3+4: same amount loaded on gel  
But concentration in 4 = 2x  
concentration in 3



11/13

Fig. 9b



12/13

Fig. 10

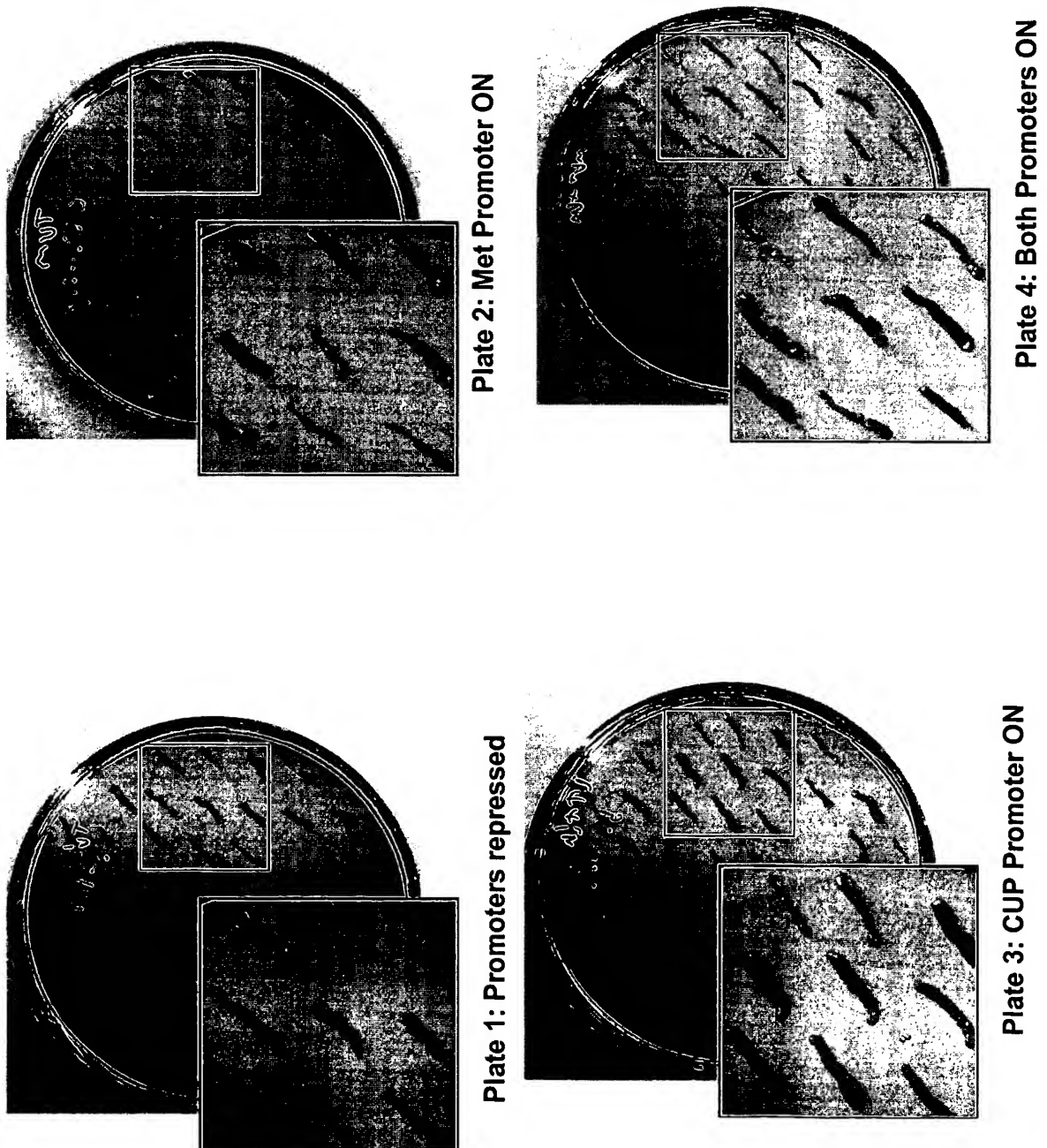
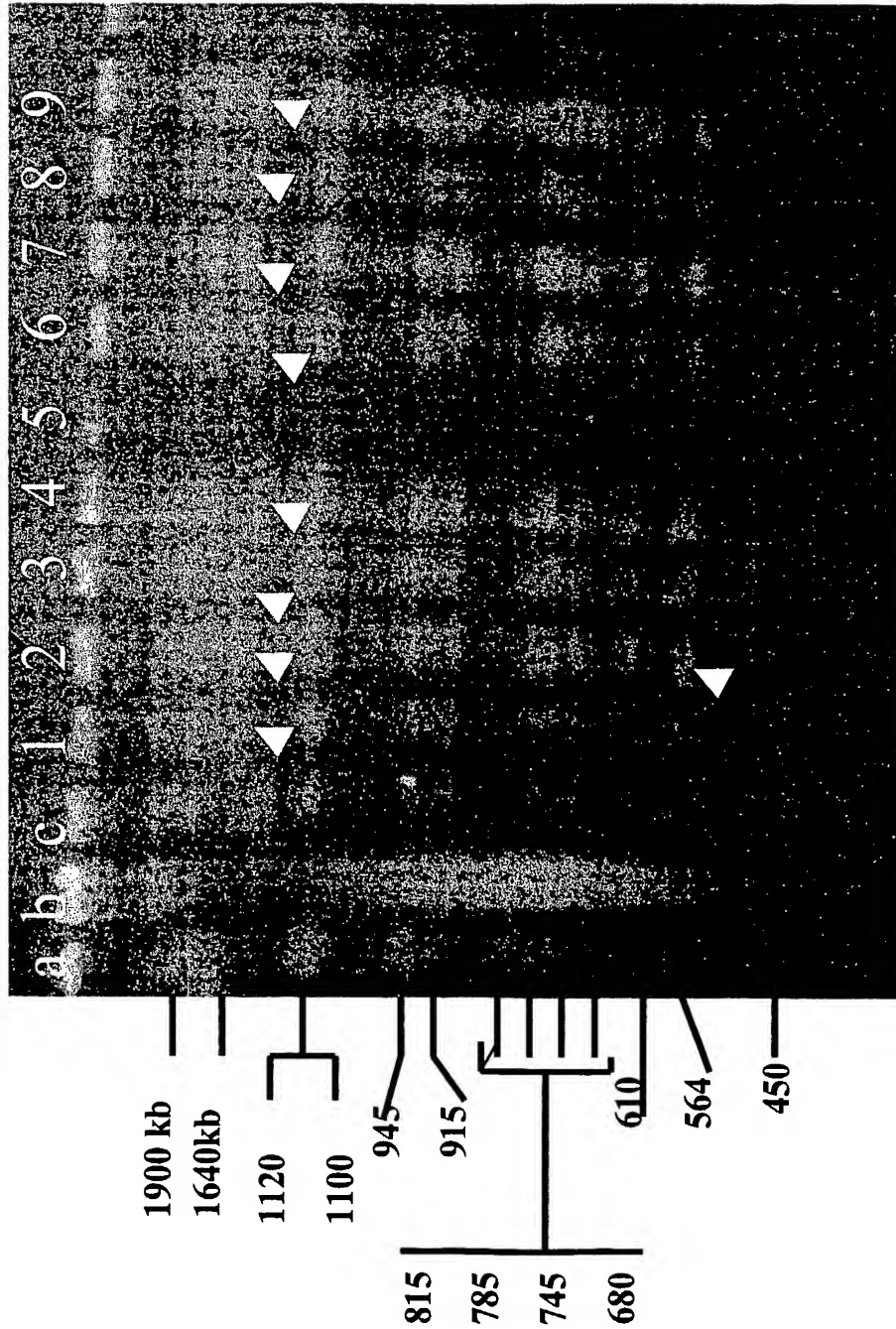


Fig. 11



## SEQUENCE LISTING

<110> Evolva Biotech AS  
Goldsmith, Neil  
Sørensen, Alexandra M. P.  
Nielsen, Søren V.S.

<120> A library of a collection of cells

<130> P 501 PC00

<150> DK PA 2001 00128  
<151> 2001-01-25

<150> DK PA 2001 00679  
<151> 2001-05-01

<150> US 60/300,863  
<151> 2001-06-27

<160> 4

<170> PatentIn version 3.1

<210> 1  
<211> 3417  
<212> DNA  
<213> Synthetic

<220>  
<221> misc\_feature  
<222> (1902)..(2759)  
<223> Ampicillin resistance gene

<220>  
<221> rep\_origin  
<222> (959)..(1899)  
<223> ColE1

<220>  
<221> misc\_feature  
<222> (2891)..(3347)  
<223> f1-phage origin of replication

<220>  
<221> terminator  
<222> (495)..(823)  
<223> ADH1

<220>  
<221> promoter  
<222> (49)..(437)  
<223> Met25 promoter

<400> 1  
 ctgatttgcc cgggcagttc aggcctcatca ggcgcgccat gcagggattc ttcggatgca 60  
 agggttcgaa tcccttagct ctcattatct tttgcttttt ctcttgaggt cacatgatcg 120  
 caaaatggca aatggcacgt gaagctgtcg atattgggga actgtggtgg ttggcaaatg 180  
 actaattaag ttagtcaagg cgccatcctc atgaaaactg tgtaacataa taaccgaagt 240  
 gtcgaaaagg tggcaccttg tccaattgaa cacgctcgat gaaaaaata agatatatat 300  
 aaggtttaagt aaagcgtctg ttagaaagga agtttttctt ttttcttgct ctcttgctct 360  
 ttcactact atttccttcg tgtaatacag ggtcgtcaga tacatagata caattctatt 420  
 acccccatcc atacaagctt ggcgccgaat tcgtcgaccc ggggatccgc ggccgcagggc 480  
 ctaaattgat ctagagcttt ggacttcttc gccagagggt tgggtcaagtc tccaatcaag 540  
 gttgtcggct tgtctacctt gccagaaatt tacgaaaaga tggaaaaggg tcaaatcggt 600  
 ggtagatacg ttgttgacac ttctaaataa gcgaatttct tatgatttat gattttttatt 660  
 attaaataag ttataaaaaa aataagtgtg tacaaaattt aaagtgactc ttaggtttta 720  
 aaacgaaaat tcttgttctt gagtaactct ttctgtagg tcagggtgct ttctcaggta 780  
 tagcatgagg tcgctcttat tgaccacacc tctaccggca tgcccatggg ttaactgatc 840  
 aatgcatcct gcatggcgcg cctgatgagc ctgaactgcc cgggcaaata agctggacgt 900  
 ctgcctgcat taatgaatcg gccaacgcgc ggggagaggc ggtttgcgta ttgggcgctc 960  
 ttccgcttcc tcgctcactg actcgctgcg ctcggtcggt cggctgcggc gagcgggtatc 1020  
 agctcactca aaggcggtta tacggttatc cacagaatca ggggataacg caggaaagaa 1080  
 catgtgagca aaaggccagc aaaaggccag gaaccgtaaa aaggccgcgt tgctggcggt 1140  
 tttccatagg ctccgcccc ctgacgagca tcacaaaaat cgacgctcaa gtcagaggtg 1200  
 gcgaaaccgc acaggactat aaagatacca ggcgtttccc cctggaagct cctcgtgcg 1260  
 ctctcctggt cgcacctgc cgcttaccgg atacctgtcc gcctttctcc ctccgggaag 1320  
 cgtggcgctt tctcatagct cacgctgtag gtatctcagt tcggtgtagg tcgttcgctc 1380  
 caagctgggc tgtgtgcacg aacccccctg tcagcccgac cgctgcgcct tatccggtaa 1440  
 ctatcgtctt gagtccaacc cggtaagaca cgacttatcg ccaactggcag cagccactgg 1500  
 taacaggatt agcagagcga ggtatgtagg cggtgctaca gagttcttga agtgggtggc 1560  
 taactacggc tacactagaa ggacagtatt tggatatctgc gctctgctga agccagttac 1620  
 ctccgaaaaa agagttggta gctcttgatc cggcaaacia accaccgctg gtagcgggtg 1680  
 tttttttggt tgcaagcagc agattacgcg cagaaaaaaa ggatctcaag aagatccttt 1740  
 gatcttttct acggggtctg acgctcagt gaacgaaaac tcacgttaag ggattttggt 1800  
 catgagatta tcaaaaagga tcttcaccta gatcctttta aattaaaaat gaagttttta 1860

atcaatctaa agtatatatg agtaaacttg gtctgacagt taccaatgct taatcagtga 1920  
ggcacctatc tcagcgatct gtctatctcg ttcattccata gttgcctgac tccccgtcgt 1980  
gtagataact acgatacggg agggccttacc atctggcccc agtgctgcaa tgataccgcg 2040  
agaccacgc tcaccggctc cagatttatc agcaataaac cagccagccg gaagggccga 2100  
gcgcagaagt ggtcctgcaa ctttatccgc ctccatccag tctattaatt gttgccggga 2160  
agctagagta agtagttcgc cagttaatag tttgcgcaac gttggtgcca ttgctacagg 2220  
catcgtggtg tcacgctcgt cgtttggtat ggcttcattc agctccggtt cccaacgac 2280  
aaggcgagtt acatgatccc ccatgttgtg caaaaaagcg gttagctcct tcgggtcctcc 2340  
gatcgttgtc agaagtaagt tggccgcagt gttatcactc atgggttatgg cagcactgca 2400  
taattctctt actgtcatgc catccgtaag atgcttttct gtgactgggtg agtactcaac 2460  
caagtcattc tgagaatagt gtatgcggcg accgagttgc tcttgcccgg cgtcaatacg 2520  
ggataatacc gcgccacata gcagaacttt aaaagtgtc atcattggaa aacgttcttc 2580  
ggggcgaaaa ctctcaagga tcttaccgct gttgagatcc agttcgatgt aaccactcg 2640  
tgcaccaaac tgatcttcag catcttttac tttcaccagc gtttctgggt gagcaaaaac 2700  
aggaaggcaa aatgccgcaa aaaagggaat aagggcgaca cggaaatgtt gaatactcat 2760  
actcttcctt tttcaatatt attgaagcat ttatcagggt tattgtctca tgagcggata 2820  
catatctgaa tgtatctaga aaaataaaca aataggggtt ccgcgcacat tccccgaaa 2880  
agtgccacct gacgcgccct gtagcggcgc attaacgcgc gcgggtgtgg tggttacgcg 2940  
cagcgtgacc gctacacttg ccagcgcctc agcgcgcgt cctttcgctt tcttccttc 3000  
ctttctcgcc acgttcgccg gctttccccg tcaagctcta aatcgggggc tccctttagg 3060  
gttccgattt agtgctttac ggcacctcga cccccaaaaa cttgattagg gtgatggttc 3120  
acgtagtggg ccacgcgccct gatagacggt ttttcgcctt ttgacgttgg agtccacgtt 3180  
ctttaatagt ggactcttgt tccaaactgg aacaacactc aaccctatct cgggtctattc 3240  
ttttgattta taagggattt tgccgatttc ggcctattgg ttaaaaaatg agctgattta 3300  
acaaaaattt aacgcgaatt ttaacaaaat attaacgctt acaatttcca ttcgccattc 3360  
aggctgcgca actgttggga agggcgatcg gtgcgggcct ctctcgtatt acgccag 3417

<210> 2  
<211> 3501  
<212> DNA  
<213> Synthetic

<220>  
<221> misc\_feature  
<222> (1986)..(2843)  
<223> Ampicillin resistance gene

~^^~

<221> rep\_origin  
 <222> (1043)..(1983)  
 <223> ColE1

<220>  
 <221> misc\_feature  
 <222> (2975)..(3431)  
 <223> fl-phage origin of replication

<220>  
 <221> terminator  
 <222> (579)..(907)  
 <223> ADH1

<220>  
 <221> promoter  
 <222> (49)..(519)  
 <223> Cup1 promoter

<400> 2  
 ctgatttgcc cgggcagttc aggctcatca ggcgcgccat gcagggataa gccgatccca 60  
 ttaccgacat ttgggcgcta tacgtgcata tgttcattgta tgtatctgta tttaaaacac 120  
 ttttgtatta tttttcctca tatatgtgta taggtttata cggatgattt aattattact 180  
 tcaccaccct ttatttcagg ctgatatctt agccttggtta ctagttagaa aaagacattt 240  
 ttgctgtcag tcactgtcaa gagattcttt tgctggcatt tcttctagaa gcaaaaagag 300  
 cgatgcgtct tttccgctga accgttcag caaaaagac taccaacgca atatggattg 360  
 tcagaatcat ataaaagaga agcaaataac tccttgtctt gtatcaattg cattataata 420  
 tcttcttggt agtgcaatat catatagaag tcatcgaaat agatattaag aaaaacaaac 480  
 tgtacaatca atcaatcaat catcacataa aatgttcaaa gcttggcgcc gaattcgtcg 540  
 acccggggat ccgcggccgc aggcctaaat tgatctagag ctttggactt cttcgccaga 600  
 ggtttggtca agtctccaat caagggtgtc ggcttgtcta ccttgccaga aatttacgaa 660  
 aagatggaaa aggttcaa atcggtgtga tacgttggtg acacttctaa ataagcgaat 720  
 ttcttatgat ttatgatttt tattattaaa taagttataa aaaaaataag tgtatacaaa 780  
 ttttaaagtg actcttaggt tttaaaacga aaattcttgt tcttgagtaa ctctttcctg 840  
 taggtcaggt tgctttctca ggtatagcat gaggtcgctc ttattgacca cacctctacc 900  
 ggcattgccc tgggttaact gatcaatgca tcctgcatgg cgcgcctgat gagcctgaac 960  
 tgccccgggca aatcagctgg acgtctgcct gcattaatga atcgccaac gcgcggggag 1020  
 aggcgggttg cgtattgggc gctcttcgc ttctcgctc actgactcgc tgcgctcggt 1080

5/16

cgttcggctg	cggcgagcgg	tatcagctca	ctcaaaaggcg	gtaatacggg	tatccacaga	1140
atcaggggat	aacgcaggaa	agaacatgtg	agcaaaaggc	cagcaaaagg	ccaggaaccg	1200
taaaaaggcc	gcgttgctgg	cgtttttcca	taggctccgc	ccccctgacg	agcatcacia	1260
aaatcgacgc	tcaagtcaga	ggtggcgaaa	cccgacagga	ctataaagat	accaggcggt	1320
tccccctgga	agctccctcg	tgcgctctcc	tggtccgacc	ctgccgctta	ccggatacct	1380
gtccgccttt	ctcccttcgg	gaagcgtggc	gctttctcat	agctcacgct	gtaggtatct	1440
cagttcgggtg	taggtcggtc	gctccaagct	gggctgtgtg	cacgaacccc	ccgttcagcc	1500
cgaccgctgc	gccttatccg	gtaactatcg	tcttgagtcc	aacccggtaa	gacacgactt	1560
atcgccactg	gcagcagcca	ctggtaacag	gattagcaga	gcgaggatatg	taggcgggtgc	1620
tacagagttc	ttgaagtggg	ggcctaacta	cggctacact	agaaggacag	tatttggtat	1680
ctgcgctctg	ctgaagccag	ttaccttcgg	aaaaagagtt	ggtagctctt	gatccggcaa	1740
acaaaccacc	gctggtagcg	gtgggtttttt	tgtttgcaag	cagcagatta	cgcgcagaaa	1800
aaaaggatct	caagaagatc	ctttgatctt	ttctacgggg	tctgacgctc	agtggaaacga	1860
aaactcacgt	taagggatctt	tggtcatgag	attatcaaaa	aggatcttca	cctagatcct	1920
tttaaattaa	aatgaagtt	ttaaatcaat	ctaaagtata	tatgagtaaa	cttgggtctga	1980
cagttaccaa	tgcttaatca	gtgaggcacc	tatctcagcg	atctgtctat	ttcggttcac	2040
catagttgcc	tgactccccg	tcgtgtagat	aactacgata	cgggagggct	taccatctgg	2100
ccccagtgct	gcaatgatac	cgcgagaccc	acgctcaccg	gctccagatt	tatcagcaat	2160
aaaccagcca	gccggaagg	ccgagcgcag	aagtggctct	gcaactttat	ccgcctccat	2220
ccagtctatt	aattgttgcc	gggaagctag	agtaagtagt	tcgccagtta	atagtttgcg	2280
caacgttggt	gccattgcta	caggcatcgt	ggtgtcacgc	tcgtcgtttg	gtatggcttc	2340
attcagctcc	ggttcccaac	gatcaaggcg	agttacatga	tcccccatgt	tgtgcaaaaa	2400
agcgggttagc	tccttcggtc	ctccgatcgt	tgtcagaagt	aagttggccg	cagtgttatc	2460
actcatgggt	atggcagcac	tgcataattc	tcttactgtc	atgccatccg	taagatgctt	2520
ttctgtgact	ggtgagtact	caaccaagtc	attctgagaa	tagtgtatgc	ggcgaccgag	2580
ttgctcttgc	ccggcgctca	tacgggataa	taccgcgcca	catagcagaa	ctttaaaagt	2640
gctcatcatt	ggaaaacggt	cttcggggcg	aaaactctca	aggatcttac	cgctgttgag	2700
atccagttcg	atgtaaccca	ctcgtgcacc	caactgatct	tcagcatctt	ttactttcac	2760
cagcgtttct	gggtgagcaa	aaacaggaag	gcaaaatgcc	gcaaaaaagg	gaataagggc	2820
gacacggaaa	tggtgaatac	tcatactctt	cctttttcaa	tattattgaa	gcatttatca	2880
gggttattgt	ctcatgagcg	gatacatatt	tgaatgtatt	tagaaaaata	aacaaatagg	2940
ggttccgcgc	acatttcccc	gaaaagtgcc	acctgacgcg	ccctgtagcg	gcgcattaag	3000



```
cgcgggcgggt gtggtgggtta cgcgccagcgt gaccgctaca cttgccagcg ccctagcgcc 3060
cgctcctttc gctttcttcc ctctctttct cgccacgttc gccggctttc cccgtcaagc 3120
tctaaatcgg gggctccctt tagggttccg atttagtgct ttacggcacc tcgaccccaa 3180
aaaacttgat taggggtgatg gttcacgtag tgggccatcg ccctgataga cggtttttcg 3240
ccctttgacg ttggagtcca cgttctttaa tagtggactc ttgttccaaa ctggaacaac 3300
actcaaccct atctcgggtct attcttttga tttataaggg attttgccga tttcggccta 3360
ttggttaaaa aatgagctga ttttaacaaa atttaacgcg aattttaaca aaatattaac 3420
gcttacaatt tccattcgcc attcaggctg cgcaactggt gggaagggcg atcggtgcg 3480
gcctcttcgc tattaagcca g 3501
```

```
<210> 3
<211> 4188
<212> DNA
<213> Synthetic
```

```
<220>
<221> misc_feature
<222> (2673)..(3530)
<223> Ampicillin resistance gene
```

```
<220>
<221> rep_origin
<222> (1730)..(2670)
<223> ColE1
```

```
<220>
<221> misc_feature
<222> (3662)..(4118)
<223> f1-phage origin of replication
```

```
<220>
<221> terminator
<222> (1027)..(1355)
<223> ADH1
```

```
<220>
<221> promoter
<222> (582)..(969)
<223> Met25 promoter
```

```
<220>
<221> misc_feature
<222> (1365)..(1603)
<223> ARS1 (autonomous replicating sequence) for Yeast replication
```

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;222&gt; (49)..(574)

&lt;223&gt; lambda spacer DNA (22428-22923)

&lt;400&gt; 3

ctgatttgcc cgggcagttc aggctcatca ggcgcgccat gcagggattc tggaaattgc	60
aacgaaggaa gaaacctcgt tgctggaagc ctggaagaag tatcgggtgt tgctgaaccg	120
tgttgataca tcaactgcac ctgatattga gtggcctgct gtccctgtta tggagtaatc	180
gttttgtgat atgccgcaga aacgttgtat gaaataacgt tctgcggtta gttagtatat	240
tgtaaagctg agtattgggt tattttggcg ttattatctt caggagaata atggaagtgc	300
tatgactcaa ttgttcatag tgtttacatc accgccatt gcttttaaga ctgaacgcat	360
gaaatatggg ttttcgtcat gttttgagtc tgctgttgat atttctaaag tcgggtttttt	420
ttcttcgttt tctctaacta ttttccatga aatacatctt tgattattat ttgaatcaat	480
tccaattacc tgaagtcttt catctataat tggcattgta tgtattgggt tattggagta	540
gatgcttgct tttctgagcc atagctctga tatcagatct tcttcggatg caagggttcg	600
aatcccttag ctctcattat tttttgcttt ttctcttgag gtcacatgat cgcaaaatgg	660
caaatggcac gtgaagctgt cgatattggg gaactgtggg ggttggcaaa tgactaatta	720
agttagtcaa ggcgccatcc tcatgaaaac tgtgtaacat aataaccgaa gtgtcgaaaa	780
ggtggcacct tgtccaattg aacacgctcg atgaaaaaaaa taagatatat ataaggttaa	840
gtaaagcgtc tgttagaaag gaagtttttc ctttttcttg ctctcttgtc ttttcatcta	900
ctatttcctt cgtgtaatac agggctgtca gatacataga tacaattcta ttaccccat	960
ccatacaagc ttggcgccga attcgtcgac ccggggatcc gcggcgcag gcctaaattg	1020
atctagagct ttggacttct tcgccagagg tttggccaag tctccaatca aggttgtcgg	1080
cttgtctacc ttgccagaaa tttacgaaaa gatggaaaag ggtcaaactg ttggtagata	1140
cgttgttgac acttctaaat aagcgaattt cttatgattt atgattttta ttattaaata	1200
agttataaaa aaaataagtg tatacaaatt ttaaagtgac tcttaggttt taaaacgaaa	1260
attcttggtc ttgagtaact ctttcctgta ggtcagggtg ctttctcagg tatagcatga	1320
ggtcgctctt attgaccaca cctctaccgg catgcccatt gggtcttttg aaaagcaagc	1380
ataaaagatc taaacataaa atctgtaaaa taacaagatg taaagataat gctaaatcat	1440
ttggctttttt gattgattgt acaggaaaat atacatcgca ggggggttgac ttttaccatt	1500
tcaccgcaat ggaatcaaac ttgttgaaga gaatgttcac aggcgcatac gctacaatga	1560
cccgattctt gctagccttt tctcggctct gcaacaacc gccaaactgat caatgcatcc	1620

tgcatggcgc	gcctgatgag	cctgaactgc	ccgggcaaat	cagctggacg	tctgcctgca	1680
ttaatgaatc	ggccaacgcg	cggggagagg	cggtttgcgt	attgggcgct	cttccgcttc	1740
ctcgctcact	gactcgctgc	gctcggctcg	tcggctgcgg	cgagcgggat	cagctcactc	1800
aaaggcggtta	atacggttat	ccacagaatc	aggggataac	gcaggaaaga	acatgtgagc	1860
aaaaggccag	caaaaggcca	ggaaccgtaa	aaaggccgcg	ttgctggcgt	ttttccatag	1920
gctccgcccc	cctgacgagc	atcacaaaaa	tcgacgctca	agtcagaggt	ggcgaaaccc	1980
gacaggacta	taaagatacc	aggcgtttcc	ccctggaagc	tccctcgctg	gctctcctgt	2040
tccgaccctg	ccgcttaccg	gatacctgtc	cgcttttctc	ccttcgggaa	gcgtggcgct	2100
ttctcatagc	tcacgctgta	ggtatctcag	ttcgggtgtag	gtcgttcgct	ccaagctggg	2160
ctgtgtgcac	gaaccccccg	ttcagcccga	ccgctgcgcc	ttatccggta	actatcgctc	2220
tgagtccaac	ccggtaagac	acgacttatc	gccactggca	gcagccactg	gtaacaggat	2280
tagcagagcg	aggtatgtag	gcggtgctac	agagttcttg	aagtgggtggc	ctaactacgg	2340
ctacactaga	aggacagtat	ttggtatctg	cgctctgctg	aagccagtta	ccttcggaaa	2400
aagagttggt	agctcttgat	ccggcaaaca	aaccaccgct	ggtagcggtg	gtttttttgt	2460
ttgcaagcag	cagattacgc	gcagaaaaaa	aggatctcaa	gaagatcctt	tgatcttttc	2520
tacggggctc	gacgctcagt	ggaacgaaaa	ctcacgttaa	gggatttttg	tcatgagatt	2580
atcaaaaagg	atcttcacct	agatcctttt	aaattaaaaa	tgaagtttta	aatcaatcta	2640
aagtatatat	gagtaaactt	ggtctgacag	ttaccaatgc	ttaatcagtg	aggcacctat	2700
ctcagcgatc	tgtctatttc	gttcacccat	agttgcctga	ctccccgtcg	tgtagataac	2760
tacgatacgg	gagggccttac	catctggccc	cagtgcctgca	atgataccgc	gagaccacgc	2820
ctcaccggct	ccagatttat	cagcaataaa	ccagccagcc	ggaagggccg	agcgcagaag	2880
tggtcctgca	actttatccg	cctccatcca	gtctattaat	tgttgccggg	aagctagagt	2940
aagtagttcg	ccagttaata	gtttgcgcaa	cgttggtgcc	attgctacag	gcatcggtgt	3000
gtcacgctcg	tcgtttggta	tggcttcatt	cagctccggt	tcccaacgat	caaggcgagt	3060
tacatgatcc	cccatgttgt	gcaaaaaagc	ggttagctcc	ttcggctcctc	cgatcgttgt	3120
cagaagtaag	ttggccgcag	tgttatcact	catggttatg	gcagcactgc	ataattctct	3180
tactgtcatg	ccatccgtaa	gatgcttttc	tgtgactggg	gagtactcaa	ccaagtcatt	3240
ctgagaatag	tgtatgcggc	gaccgagttg	ctcttgcccc	gcgtcaatac	gggataatac	3300
cgcgccacat	agcagaactt	taaaagtgc	catcattgga	aaacgttctt	cggggcgaaa	3360
actctcaagg	atcttaccgc	tggtgagatc	cagttcgatg	taaccactc	gtgcacccaa	3420
ctgatcttca	gcatctttta	ctttcaccag	cgtttctggg	tgagcaaaaa	caggaaggca	3480
aaatgccgca	aaaaagggaa	taagggcgac	acggaaatgt	tgaatactca	tactcttcct	3540

```
ttttcaatat tattgaagca tttatcaggg ttattgtctc atgagcggat acatatttga 3600
atgtatttag aaaaataaac aaataggggt tccgcgcaca tttccccgaa aagtgccacc 3660
tgacgcgccc tgtagcggcg cattaagcgc ggcggggtgtg gtgggttacgc gcagcgtgac 3720
cgctacactt gccagcgcgc tagcgcgcgc tcctttcgtt ttcttccctt cctttctcgc 3780
cacgttcgcc ggctttcccc gtcaagctct aaatcggggg ctcccttttag ggttccgatt 3840
tagtgcttta cggcacctcg accccaaaaa acttgattag ggtgatgggt cacgtagtgg 3900
gccatcgccc tgatagacgg tttttcgcgc ttgacgttg gagtccacgt tctttaatat 3960
tggaactcttg ttccaaactg gaacaacact caaccctatc tcggtctatt cttttgattt 4020
ataagggatt ttgccgattt cggcctattg gttaaaaaat gagctgattt aacaaaaatt 4080
taacgcgaat ttttaacaaa tattaacgct tacaatttcc attcgccatt caggctgcgc 4140
aactgttggg aagggcgatc ggtgcggggc tcttcgctat tacgccag 4188
```

```
<210> 4
<211> 11466
<212> DNA
<213> Synthetic
```

```
<220>
<221> misc_feature
<222> (3560)..(4247)
<223> Tetrahymena thermophila macronuclear telomere
```

```
<220>
<221> misc_feature
<222> (6024)..(6711)
<223> Tetrahymena thermophila macronuclear telomere
```

```
<220>
<221> misc_feature
<222> (9644)..(10388)
<223> Autonomous replicating sequence
```

```
<220>
<221> misc_feature
<222> (10488)..(11465)
<223> Centromere IV
```

```
<220>
<221> rep_origin
<222> (7198)..(7198)
<223> Origin of replication, PMB1
```

```
<220>
<221> misc_feature
<222> (1962)..(2765)
<223> URA3, orotidine-5'-phosphate decarboxylase coding sequence
```

```

<220>
<221> misc_feature
<222> (4893)..(5552)
<223> HIS3, imidazoleglycerolphosphate dehydratase, coding sequence

<220>
<221> misc_feature
<222> (7956)..(8816)
<223> AP(R), beta-lactamase, ampR ampicillin resistance, coding sequence

<220>
<221> misc_feature
<222> (9129)..(9803)
<223> TRP1, phosphoribosylanthranilate isomerase, coding sequence

<400> 4
ttctcatggt tgacagctta tcatcgataa gctttaatgc ggtagtttat cacagttaaa      60
ttgctaacgc agtcaggcac cgtgtatgaa atctaacaat gcgctcatcg tcatcctcgg      120
caccgtcacc ctggatgctg taggcatagg cttggttatg ccggtactgc cgggcctcct      180
gcgggatatc gtccattccg acagcatcgc cagtcactat ggcgtgctgc tagcgctata      240
tgcgttgatg caatttctat gcgcacccgt tctcggagca ctgtccgacc gctttggccg      300
ccgcccagtc ctgctcgctt cgctacttgg agccactatc gactacgcca tcatggcgac      360
cacaccgctc ctgtggatca attcccttta gtataaattt cactctgaac catcttggaa      420
ggaccggtaa ttatttcaaa tctctttttc aattgtatat gtgttatggt atgtagtata      480
ctctttcttc aacaattaaa tactctcggg agccaagttg gtttaaggcg caagacttta      540
atztatcact acggaattgg cgcgccaatt ccgtaatctt gagatcgggc gttcgatcgc      600
cccgggagat ttttttgttt tttatgtctt ccattcactt ccagacttg caagttgaaa      660
tatttctttc aagggaattg atcctctacg ccggacgcat cgtggccggc atcaccggcg      720
ccacaggtgc ggttgctggc gcctatatcg ccgacatcac cgatggggaa gatcgggctc      780
gccacttcgg gctcatgagc gcttgtttcg gcgtgggtat ggtggcaggc cccgtggccg      840
ggggactggt gggcgccatc tccttgcagc caccattcct tgcggcgggc gtgctcaacg      900
gcctcaacct actactgggc tgcttcctaa tgcaggagtc gcataaggga gagcgtcgac      960
cgatgccctt gagagccttc aaccagtcga gctccttcgg gtgggcgcgg ggcattgacta     1020
tcgtcgccgc acttatgact gtcttcttta tcatgcaact cgtaggacag gtgccggcag     1080
cgctctgggt cattttcggc gaggaccgct ttcgctggag cgcgacgatg atcggcctgt     1140
cgcttgcggt attcggaatc ttgcacgccc tcgctcaagc cttegtcact ggtcccgcga     1200
ccaaacgttt cggcgagaag caggccatta tcgcgggcat ggcggccgac gcgctgggct     1260

```

acgtcttgct	ggcgttcgcg	acgcgaggct	ggatggcctt	ccccattatg	attctttctcg	1320
cttccggcgg	catcgggatg	cccgcgttgc	aggccatgct	gtccaggcag	gtagatgacg	1380
accatcaggg	acagcttcaa	ggatcgctcg	cggctcttac	cagcctaact	tcgatcactg	1440
gaccgctgat	cgtcacggcg	atztatgccg	cctcggcgag	cacatggaac	gggttggcat	1500
ggattgtagg	cgccgcccta	taccttgtct	gcctccccgc	gttgcgtcgc	ggatgcattg	1560
gccggggcac	ctcgacctga	atggaagccg	gcggcacctc	gctaacggat	tcaccactcc	1620
aagaattgga	gccaatcaat	tcttgcgagg	aactgtgaat	gcgcaaacca	acccttggca	1680
gaacatatcc	atcgcgcccg	ccatctccag	cagccgcacg	cggcgcatcc	ccccccccct	1740
ttcaattcaa	ttcatcattt	tttttttatt	cttttttttg	atttcgggtt	ctttgaaatt	1800
tttttgattc	ggtaatctcc	gaacagaagg	aagaacgaag	gaaggagcac	agacttagat	1860
tggatatatat	acgcatatgt	agtgttgaag	aaacatgaaa	ttgccagta	ttcttaaccc	1920
aactgcacag	aacaaaaacc	tgcaggaaac	gaagataaat	catgtcgaaa	gctacatata	1980
aggaacgtgc	tgctactcat	cctagtcctg	ttgctgcaa	gctatttaat	atcatgcacg	2040
aaaagcaaac	aaacttgtgt	gcttcattgg	atgttcgtac	caccaaggaa	ttactggagt	2100
tagttgaagc	attaggtccc	aaaatttggt	tactaaaaac	acatgtggat	atcttgactg	2160
atttttccat	ggagggcaca	gttaagccgc	taaaggcatt	atccgccaa	tacaattttt	2220
tactcttcga	agacagaaaa	tttgctgaca	ttggtaatac	agtcaaattg	cagtactctg	2280
cgggtgtata	cagaatagca	gaatgggcag	acattacgaa	tgcacacggg	gtggtggggc	2340
caggtattgt	tagcggtttg	aagcaggcgg	cagaagaagt	aacaaaggaa	cctagaggcc	2400
ttttgatggt	agcagaattg	tcattgcaag	gctccctatc	tactggagaa	tatactaagg	2460
gtactgttga	cattgcgaag	agcgacaaa	attttggttat	cggctttatt	gctcaaagag	2520
acatgggtgg	aagagatgaa	ggttacgatt	ggttgattat	gacacccggg	gtgggttttag	2580
atgacaaggg	agacgcattg	ggtcaacagt	atagaaccgt	ggatgatgtg	gtctctacag	2640
gatctgacat	tattattggt	ggaagaggac	tatttgcaaa	gggaagggat	gctaaggtag	2700
aggggtgaac	ttacagaaaa	gcaggctggg	aagcatattt	gagaagatgc	ggccagcaaa	2760
actaaaaaac	tgtattataa	gtaaatgcat	gtatactaaa	ctcaciaaatt	agagcttcaa	2820
tttaattata	tcagttatta	ctcgggcgta	atgattttta	taatgacgaa	aaaaaaaaaa	2880
ttggaaagaa	aagggggggg	gggcagcggt	gggtcctggc	cacgggtgcg	catgatcgtg	2940
ctcctgtcgt	tgaggacccg	gctaggctgg	cgggggttgc	ttactgggta	gcagaatgaa	3000
tcaccgatac	gcgagcgaac	gtgaagcgac	tgctgctgca	aaacgtctgc	gacctgagca	3060
acaacatgaa	tggtcttcgg	tttccgtggt	tcgtaaagtc	tggaaacgcg	gaagtcagcg	3120

ccctgcacca	ttatgttccg	gatctgcac	gcaggatgct	gctggctacc	ctgtggaaca	3180
cctacatctg	tattaacgaa	gcgctggcat	tgaccctgag	tgatttttct	ctgggtcccg	3240
cgcattccata	ccgccagttg	tttaccctca	caacgttcca	gtaaccgggc	atgttcatca	3300
tcagtaaccc	gtatcgtgag	catcctctct	cgtttcatcg	gtatcattac	ccccatgaac	3360
agaaattccc	ccttacacgg	aggcatcaag	tgaccaaaca	ggaaaaaacc	gcccttaaca	3420
tggcccgctt	tatcagaagc	cagacattaa	cgcttctgga	gaaactcaac	gagctggacg	3480
cggatgaaca	ggcagacatc	tgtgaatcgc	ttcacgacca	cgctgatgag	ctttaccgca	3540
gccctcgagg	gataagcttc	attttttagat	aaaatttatt	aatcatcatt	aatttcttga	3600
aaaacatttt	atttattgat	cttttataac	aaaaaacctt	tctaaaagtt	tatttttgaa	3660
tgaaaaactt	ataaaaaattt	atgaaaacta	caaaaaataa	aatttttaat	taaaataatt	3720
ttgataagaa	cttcaatctt	tgactagcta	gcttagtcat	ttttgagatt	taattaatat	3780
tttatgttta	ttcatatata	aactattcaa	aatattatag	aatttaaaca	ttttaacatc	3840
ttaatcattc	ataaataact	aaaaatcaaa	gtattacatc	aataaataac	ttttactcaa	3900
tgtcaaagaa	ttattggggg	tgggggttgg	gttgggggtt	gggttggggg	tgggggttgg	3960
gttgggggtt	gggttggggg	tgggggttgg	gttgggggtt	gggttggggg	tgggggttgg	4020
gttgggggtt	gggttggggg	tgggggttgg	gttgggggtt	gggttggggg	tgggggttgg	4080
gttgggggtt	gggttggggg	tgggggttgg	gttgggggtt	gggttggggg	tgggggttgg	4140
gttgggggtt	gggttggggg	tgggggttgg	gttgggggtt	gggttggggg	acagcattca	4200
gggtattagaa	gaatatcctg	attcaggtga	aaatattgtt	gatgcgcggg	atcctcgggg	4260
acaccaaata	tggcgatctc	ggccttttct	tttcttggag	ctgggacatg	tttgccatcg	4320
atccatctac	caccagaacg	gccgttagat	ctgctgccac	cgttgtttcc	accgaagaaa	4380
ccaccgttgc	cgtaaccacc	acgacggttg	ttgctaaaga	agctgccacc	gccacggcca	4440
ccgttgtagc	cgcggttggt	gttattgtag	ttgctcatgt	tatttctggc	acttcttggt	4500
tttctcttta	agtgaggagg	aacataacca	ttctcgttgt	tgctcgtgat	gcttaaattt	4560
tgcacttggt	cgctcagttc	agccataata	tgaaatgctt	ttcttggtgt	tcttacggaa	4620
taccacttgc	cacctatcac	cacaactaac	tttttcccgt	tcctccatct	cttttatatt	4680
ttttttctcg	atcgagttca	agagaaaaaa	aaagaaaaag	caaaaagaaa	aaaggaaagc	4740
gcgcctcggt	cagaatgaca	cgtatagaat	gatgcattac	cttgctcatct	tcagtatcat	4800
actgttcgta	tacatactta	ctgacattca	taggtataca	tatatacaca	tgtatatata	4860
tcgtatgctg	cagcttttaa	taatcggtgt	cactacataa	gaacaccttt	gggtggaggga	4920
acatcggttg	taccattggg	cgagggtggc	tctcttatgg	caaccgcaag	agccttgaac	4980
gcactctcac	tacgggtgat	atcattcttg	cctcgcagac	aatcaacgtg	gagggtaatt	5040

ctgctagcct	ctgcaaagct	ttcaagaaaa	tgcgggatca	tctcgcaaga	gagatctcct	5100
actttctccc	tttgcaaacc	aagttcgaca	actgcgtacg	gcctgttcga	aagatctacc	5160
accgctctgg	aaagtgcctc	atccaaaggc	gcaaatcctg	atccaaacct	ttttactcca	5220
cgcgccagta	gggcctcttt	aaaagcttga	ccgagagcaa	tcccgcagtc	ttcagtgggtg	5280
tgatggtcgt	ctatgtgtaa	gtcaccaatg	cactcaacga	ttagcgacca	gccggaatgc	5340
ttggccagag	catgtatcat	atgggtccaga	aaccctatac	ctgtgtggac	gttaatcact	5400
tgcgattgtg	tggcctgttc	tgtacttgct	tctgcctctt	tttctgggaa	gatcgagtgc	5460
tctatcgcta	ggggaccacc	ctttaagag	atcgcaatct	gaatcttggg	ttcattttgta	5520
atacgcttta	ctagggtctt	ctgctctgtc	atctttgcct	tcgtttatct	tgcttgctca	5580
ttttttagta	tattcttcga	agaaatcaca	ttactttata	taatgtataa	ttcattatgt	5640
gataatgcca	atcgctaaga	aaaaaaaaaga	gtcatccgct	aggtggaaaa	aaaaaaaaatga	5700
aatcattac	cgaggcataa	aaaaatatag	agtgtactag	aggaggccaa	gagtaataga	5760
aaaagaaaat	tgcgggaaag	gactgtgtta	tgacttcctt	gactaatgcc	gtgttcaaac	5820
gatacctggc	agtgactcct	agcgctcacc	aagctcttaa	aacgagaatt	aagaaaaagt	5880
cgtcatcttt	cgataagttt	ttcccacagc	aaagcaatag	tagaaaaaaa	caatgggaaa	5940
cgttgaatga	agacaaagcg	tcgtgggtta	aaaggaaata	cgctcacgta	catgctaggg	6000
aacaggaccg	tgcagcggat	cccgcgcata	aacaatatct	tcacctgaat	caggatatct	6060
ttctaatacc	tgaatgctgt	ttcccacccc	caaccccaac	cccaacccca	accccaaccc	6120
caaccccaac	cccaacccca	accccaaccc	caaccccaac	cccaacccca	accccaaccc	6180
caaccccaac	cccaacccca	accccaaccc	caaccccaac	cccaacccca	accccaaccc	6240
caaccccaac	cccaacccca	accccaaccc	caaccccaac	cccaacccca	accccaaccc	6300
caaccccaac	cccaacccca	accccaaccc	caaccccaac	cccaacccca	accccaataa	6360
ttctttgaca	ttgagtaaaa	gttatcttatt	gatgtaatac	tttgattttt	agttattttat	6420
gaatgattaa	gatgttaaaa	tgttttaaatt	ctataatatt	ttgaatagtt	tatatatgaa	6480
taaacataaa	atattaatta	aatctcaaaa	atgactaagc	tagctagtca	aagattgaag	6540
ttcttatcaa	aattatttta	attaaaaatt	ttatTTTTTg	tagttttcat	aaatttttat	6600
aagtttttca	ttcaaaaata	aactttttaga	agggtttttt	gttataaaaag	atcaataaat	6660
aaaatgtttt	tcaagaaatt	aatgatgatt	aataaatttt	atctaaaaat	gaagcttatc	6720
cctcgagggc	tgcctcgcgc	gtttcgggtga	tgacgggtgaa	aacctctgac	acatgcagct	6780
cccggagacg	gtcacagctt	gtctgtaagc	ggatgccggg	agcagacaag	cccgtcaggg	6840
cgcgctcagc	gggtgtggcg	gggtgcgggg	cgcagccatg	accagtcac	gtagcgatag	6900



cggagtgtat	actggcttaa	ctatgcggca	tcagagcaga	ttgtactgag	agtgcaccat	6960
atgcggtgtg	aaataccgca	cagatgcgta	aggagaaaat	accgcatcag	gcgctcttcc	7020
gcttcctcgc	tcactgactc	gctgcgctcg	gtcgttcggc	tgcggcgagc	ggatcacgct	7080
cactcaaagg	cggtaatacg	gttatccaca	gaatcagggg	ataacgcagg	aaagaacatg	7140
tgagcaaaag	gccagcaaaa	ggccaggaac	cgtaaaaagg	ccgcgttgct	ggcgtttttc	7200
cataggctcc	gccccctga	cgagcatcac	aaaaatcgac	gctcaagtca	gaggtggcga	7260
aaccgcacag	gactataaag	ataccaggcg	tttccccctg	gaagctccct	cgtgcgctct	7320
cctgttccga	ccctgccgct	taccggatac	ctgtccgcct	ttctcccttc	gggaagcggtg	7380
gcgctttctc	atagctcacg	ctgtagggtat	ctcagttcgg	tgtaggtcgt	tcgctccaag	7440
ctgggctgtg	tgacgaacc	ccccgttcag	cccgaccgct	gcgccttata	cggtaaactat	7500
cgtcttgagt	ccaacccggt	aagacacgac	ttatcgccac	tggcagcagc	cactggtaac	7560
aggattagca	gagcgaggta	tgtaggcggt	gctacagagt	tcttgaagtg	gtggcctaac	7620
tacggctaca	ctagaaggac	agtatttggt	atctgcgctc	tgctgaagcc	agttaccttc	7680
ggaaaaagag	ttggtagctc	ttgatccggc	aaacaaacca	ccgctggtag	cggtggtttt	7740
tttgtttgca	agcagcagat	tacgcgcaga	aaaaaaggat	ctcaagaaga	tcctttgatc	7800
ttttctacgg	ggctctgacgc	tcagtggaac	gaaaactcac	gttaagggat	tttggtcatg	7860
agattatcaa	aaaggatctt	cacctagatc	cttttaaatt	aaaaatgaag	ttttaaatca	7920
atctaaagta	tatatgagta	aacttgggtc	gacagttacc	aatgcttaat	cagtgaggca	7980
cctatctcag	cgatctgtct	atttcgttca	tccatagttg	cctgactccc	cgctcgtag	8040
ataactacga	tacgggaggg	cttaccatct	ggccccagtg	ctgcaatgat	accgcgagac	8100
ccacgctcac	cggctccaga	tttatcagca	ataaaccagc	cagccggaag	ggccgagcgc	8160
agaagtggtc	ctgcaacttt	atccgcctcc	atccagtcta	ttaattggtg	ccgggaagct	8220
agagtaagta	gttcgccagt	taatagtttg	cgcaacgttg	ttgccattgc	tgccagcatc	8280
gtgggtgtcac	gctcgtcggt	tggtatggct	tcattcagct	ccggttccca	acgatcaagg	8340
cgagttacat	gatcccccat	gttggtgcaa	aaagcggtta	gctccttcgg	tcctccgatc	8400
gttgctcagaa	gtaagttggc	cgcagtgtta	tcactcatgg	ttatggcagc	actgcataat	8460
tctcttactg	tcatgccatc	cgtaagatgc	ttttctgtga	ctgggtgagta	ctcaaccaag	8520
tcattctgag	aatagtgtat	gcggcgaccg	agttgctctt	gcccggcgtc	aacacgggat	8580
aataccgcgc	cacatagcag	aactttaaaa	gtgctcatca	ttggaaaacg	ttcttcgggg	8640
cgaaaactct	caaggatctt	accgctgttg	agatccagtt	cgatgtaacc	cactcgtgca	8700
cccaactgat	cttcagcatc	ttttactttc	accagcgttt	ctgggtgagc	aaaaacagga	8760
aggcaaaatg	ccgcaaaaaa	gggaataagg	gcgacacgga	aatggtgaat	actcatactc	8820

ttccttttttc	aatattatttg	aagcattttat	caggggttatt	gtctcatgag	cggatacata	8880
tttgaatgta	tttagaaaaa	taaacaaata	gggggttcgc	gcacatttcc	ccgaaaagtg	8940
ccacctgacg	tctaagaaac	cattattatc	atgacattaa	cctataaaaa	taggcgtatc	9000
acgaggccct	ttcgtcttca	agaattaatt	cggtcgaaaa	aagaaaagga	gagggccaag	9060
agggagggca	ttggtgacta	ttgagcacgt	gagtatacgt	gattaagcac	acaaaggcag	9120
cttggagtat	gtctgttatt	aatttcacag	gtagttctgg	tccattggtg	aaagtttgcg	9180
gcttgcagag	cacagaggcc	gcagaatgtg	ctctagattc	cgatgctgac	ttgctgggta	9240
ttatatgtgt	gccaataga	aagagaacaa	ttgaccgggt	tattgcaagg	aaaatttcaa	9300
gtcttgtaaa	agcatataaa	aatagttcag	gcactccgaa	atacttggtt	ggcgtgtttc	9360
gtaatcaacc	taaggaggat	gttttggctc	tgggtcaatga	ttacggcatt	gatatcgtcc	9420
aactgcatgg	agatgagtcg	tggcaagaat	accaagagtt	cctcggtttg	ccagttatta	9480
aaagactcgt	atttccaaaa	gactgcaaca	tactactcag	tgcagcttca	cagaaacctc	9540
attcgtttat	tcccttgttt	gattcagaag	caggtgggac	aggtgaactt	ttggattgga	9600
actcgatttc	tgactgggtt	ggaaggcaag	agagccccga	aagcttacat	tttatgttag	9660
ctgggtggact	gacgccagaa	aatgttgggtg	atgcgcttag	attaaatggc	gttattgggtg	9720
ttgatgtaag	cggagggtgtg	gagacaaatg	gtgtaaaaga	ctctaacaaa	atagcaaatt	9780
tcgtcaaaaa	tgctaagaaa	taggttatta	ctgagtagta	tttatttaag	tattgtttgt	9840
gcacttgcct	gcaggccctt	tgaagagcaa	gcataaaaga	tctaaacata	aaatctgtaa	9900
aataacaaga	tgtaaagata	atgctaaatc	atttggcttt	ttgattgatt	gtacaggaaa	9960
atatacatcg	caggggggttg	actttttacca	tttcaccgca	atggaatcaa	acttggttgaa	10020
gagaatgttc	acaggcgcac	acgctacaat	gacccgattc	ttgctagcct	tttctcggtc	10080
ttgcaaacaa	ccgccggcag	cttagtatat	aaatacacat	gtacatacct	ctctccgtat	10140
cctcgtaatc	attttcttgt	atttatcgtc	ttttcgctgt	aaaaacttta	tcacacttat	10200
ctcaaataca	cttattaacc	gcttttacta	ttatcttcta	cgctgacagt	aatatcaaac	10260
agtgcacacat	attaaacaca	gtgggtttctt	tgcataaaca	ccatcagcct	caagtcgtca	10320
agtaaagatt	tcgtgttcat	gcagatagat	aacaatctat	atgttgataa	ttagcgttgc	10380
ctcatcaatg	cgagatccgt	ttaaccggac	cctagtgcac	ttaccccacg	ttcgggtccac	10440
tgtgtgccga	acatgctcct	tcactatttt	aacatgtgga	attaattcta	aatcctcttt	10500
atatgatctg	ccgatagata	gttctaagtc	attgaggttc	atcaacaatt	ggattttctg	10560
tttactcgac	ttcaggtaaa	tgaaatgaga	tgatacttgc	ttatctcata	gttaactcta	10620
agaggtgata	cttattttact	gtaaaactgt	gacgataaaa	ccggaaggaa	gaataagaaa	10680

```
actcgaactg atctataatg cctatTTTTct gtaaagagtt taagctatga aagcctcggc 10740
atTTtgGCCg ctCctaggta gtGctTTTTt tccaaggaca aaacagtttc ttttcttga 10800
gcaggTTTTa tgtttcggta atcataaaca ataaataaat ttttcatTT atgtttaaaa 10860
ataaaaaata aaaaagtatt ttaaatTTTT aaaaaagttg attataagca tgtgaccttt 10920
tgcaagcaat taaatTTTgc aatttgTgat tttaggcaaa agttacaatt tctggctcgt 10980
gtaatatatg tatgctaaag tgaactTTta caaagtcgat atggacttag tcaaaagaaa 11040
tttctttaa aatatatagc actagccaat ttagcacttc tttatgagat atattataga 11100
ctttattaag ccagattTgt gtatttatatg tatttaccCG gcgaatcatg gacatacatt 11160
ctgaaatagg taatattctc tatggtgaga cagcatagat aacctaggat acaagttaa 11220
agctagtact gttttgcagt aatTTTTTtc ttttttataa gaatgttacc acctaaataa 11280
gttataaagt caatagttaa gtttgatatt tgattgtaaa ataccgtaat atatttgcat 11340
gatcaaaagg ctcaatgttg actagccagc atgtcaacca ctatattgat caccgatata 11400
tggaactcca caccaactag taatatgaca ataaattcaa gatattcttc atgagaatgg 11460
cccaga 11466
```